Data driven methods to support decision making in Deep Brain Stimulation for Parkinson's disease

Maxime Péralta

UNIVERSITÉ DE RENNES 1

THÈSE DE DOCTORAT DE

L'UNIVERSITÉ DE RENNES 1

ÉCOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : Signal, Image, Vision

Par

« Maxime PÉRALTA »

« Data driven methods to support decision making in Deep Brain Stimulation for Parkinson's Disease »

Thèse présentée et soutenue à « Rennes », le « 3 Novembre 2020 »
Unité de recherche : MediCIS - Inserm - UMR 1099 LTSI

Rapporteurs avant soutenance :

Benoît Dawant	Professeur des Universités, Vanderbilt University, Nashville, USA
Sébastien Ourselin	Professeur des Universités, King's College, London, GB

Composition du Jury :

Président :	Carine Karachi	Professeur des Universités, Sorbonne Université, Paris, France
Examinateurs :	Pierre Jannin	Directeur de Recherche INSERM, Université de Rennes 1, France
	Claire Haegelen	Professeur des Universités, Université de Rennes 1, France
	Julie Péron	Maître-assistante à l'Université de Genève, Université de Genève, Suisse
	Andréas Horn	Chercheur à la Charité Universitätsmedizin, Berlin, Allemagne
Dir. de thèse :	Pierre Jannin	Directeur de Recherche INSERM, Université de Rennes 1, France
Co-dir. de thèse :	Claire Haegelen	Professeur des Universités, Université de Rennes 1, France

Acknowledgement

Mes premiers remerciements vont à Pierre et à Claire, pour m'avoir fait très vite confiance, pour leur encadrement et pour leurs encouragements, ainsi que pour le temps précieux qu'ils m'ont consacré. Je tiens à remercier chaleureusement John, qui m'a appris tout ce que je devais savoir sur la recherche, sur le machine learning, sur la rédaction d'articles et bien d'autres. Son implication a été plus que déterminante à presque tous les niveaux de ma thèse, et si je suis aujourd'hui docteur c'est en grande partie grâce à lui.

Je tiens bien évidemment à remercier mes rapporteurs de thèse, les professeurs Benoît Dawant et Sébastien Ourselin, ainsi que mon jury, la professeure Carine Karachi et les docteurs Andréas Horn et Julie Péron. Le temps qu'ils ont consacré pour ma thèse a été considérable, et leurs retours sur mon travail précieux et encourageants. De la même manière, je voudrais remercier mon comité de suivi individuel de thèse, à savoir le professeur Marc Vérin et les docteurs Oscar Acosta et Sophie Langouët-Prigent. Leurs retours et leurs encouragements lors de mes deux premières années de thèse m'ont sincèrement aidé et remotivé.

Un grand merci à la Fondation pour la Recherche Médicale (FRM) de nous avoir fait confiance en finançant notamment ma thèse, je suis dans la sincère espérance que mon travail aura été à la hauteur de leurs attentes et de l'investissement qui a été fait.

De nombreux événements marquants ont composé ma thèse. En premier lieu je voudrais remercier le docteur Ali R. Khan de m'avoir accueilli dans son équipe au Canada, son implication et sa confiance m'auront permis de publier mon premier article dans un journal scientifique. Je voudrais encore le remercier lui et son équipe, et en particulier Dimuthu, ainsi que Yolanda, Adriana, Kelvin, Paris et Irina, pour m'avoir fait un excellent accueil et avoir fait de ces trois mois un de mes meilleurs souvenirs. Durant cette thèse, j'ai eu aussi l'occasion de m'essayer à l'enseignement, en encadrant des travaux pratiques d'abord à la Faculté des Sciences Economiques de Rennes 1, puis à l'ENSAI. Je voudrais de ce fait remercier Romaric pour toute sa confiance et pour m'avoir offert cette opportunité. Enfin, durant cette thèse, j'ai eu l'occasion d'encadrer trois stagiaires, Antoine, Anh et Alicia. Ce fut un réel plaisir et un challenge pour moi,

et j'espère sincèrement que j'aurais été à la hauteur de cette responsabilité qui me tenait à coeur, et j'espère que vous garderez un bon souvenir de votre passage au labo.

Je voudrais remercier tous mes collègues. Alfonso, Alicia, Anh, Arnaud, Arthur, Bernard, Boubacar, Clément, Delphine, Duygu, Ehouarn, Fabien, Gaetan, Gurvan, John, Julien, Kévin, Lise, Maela, Marie (merci pour la charte graphique !), Marie-Stéphanie, Marine, Naoyo, Noémie, Olga, Thibaut, Thibault, Tristan, Zakaria. Sans vous, votre bonne humeur et votre soutien, je n'aurais certainement pas été capable de tenir jusqu'au bout. Vous m'avez tous, sans exception et à votre manière, marqué, influencé et apporté quelque chose.

Ces remerciements vont également à mes amis proches, Marie, Daly, Florian, Marie et bien d'autres, ainsi qu'à ma famille, toutes ces personnes qui m'ont accompagné et qui ont su me soutenir durant ces trois années, même dans les moments de doutes et d'échec. Les résultats que j'ai obtenus et mon humeur étaient fortement interdépendants, et je n'aurais pas été capable d'aller bien loin sans vous.

Enfin, je voudrais remercier Alexandra Elbakyan pour rendre la recherche accessible à tous.

Table of Contents

Résumé étendu

La maladie de Parkinson est une maladie neurodégénérative dite des mouvements anormaux qui touche environ 1% de la population âgée de plus de 60 ans. C'est une maladie dont on ne guérit pas. Cependant, plusieurs traitements existent pour en limiter les effets moteurs et ainsi soulager le patient et améliorer sa qualité de vie. Parmi ces traitements existe la stimulation cérébrale profonde (SCP). La SCP consiste à implanter une ou deux électrodes dans le cerveau du patient afin d'y stimuler électriquement des structures sous-corticales précises, tel que le noyau sous-thalamique. Prise en complément de traitements médicamenteux, cette stimulation donne des effets spectaculaires sur la motricité du patient. Malheureusement, cette procédure comporte des risques, comme toute chirurgie, et peut aussi engendrer des effets secondaires. En outre, elle pose de nombreux challenges aux cliniciens tout au long du circuit clinique du patient.

Dans cette thèse, nous avons commencé en introduction par dresser un état des lieux de cette procédure, et nous avons isolé plusieurs défis cliniques, leurs enjeux et les problématiques techniques qui en découlent. Une difficulté supplémentaire est que la connaissance autour de la progression des maladies, ainsi que les mécanismes de la stimulation elle-même demeure limitée, et difficilement élucidable. Pour cette raison, et en vertu de la collecte de données extensive effectuée tout au long de la procédure, de nombreuses équipes de recherche ont essayé d'utiliser l'apprentissage machine, qui est un champ de recherche entre statistiques et intelligence artificielle. Les deux objectifs principaux sont de proposer, de manière prospective, des outils d'assistance à la prise de décision ou d'apporter de la connaissance par l'analyse automatique de données cliniques rétrospectives.

Nous avons, dans une deuxième partie d'introduction, défini et expliqué les principes fondamentaux de l'apprentissage machine, ses limites ainsi que les raisons techniques pour lesquelles cet outil peut théoriquement être intéressant dans le cadre de la stimulation cérébrale profonde.

Nous avons poursuivi cette introduction en réalisant une revue systématique de la littérature centrée sur l'apprentissage machine pour la stimulation cérébrale profonde. Nous avons isolé un corpus de 55 publications, que nous avons ensuite analysées en fonction de plusieurs dimensions. Plusieurs conclusions se sont dégagées de cette revue, notamment

une qui s'avérera cruciale pour la suite de cette thèse: l'analyse par approche pilotée par les données. En effet, nous avons pu distinguer deux sous-groupes d'études d'apprentissage automatique: les approches d'ingénierie basées sur des caractéristiques, et les approches pilotées par les données. Cette première approche, largement majoritaire dans notre revue, utilise les connaissances et les hypothèses humaines (techniques et cliniques) afin de transformer les données brutes sous une forme réduite, synthétique. Cette approche est intéressante, car elle rend la tâche du modèle statistique, ou d'apprentissage machine plus aisée: plus les entrées du modèle sont limitées en nombre, et plus elles sont directement corrélées à la sortie à prédire, plus l'entraînement du modèle aura de chances de réussir et de donner de bons résultats. Cependant, cette approche expose grandement les performances du système à la connaissance humaine. En effet, le modèle d'apprentissage machine ne voit en entrée que ce que l'ingénieur ou le chercheur a décidé de lui montrer, le privant ainsi de potentielles informations cruciales pour réaliser sa tâche. Afin de s'affranchir de la nécessité d'intellectualisation du problème par l'humain, une approche pilotée par les données peut être préférée. Cette approche consiste à fournir au modèle l'information sous sa forme la plus brute possible, et lui confie la tâche d'en distinguer et d'en extraire automatiquement les caractéristiques les plus pertinentes. Bien que plus ardus à entraîner, les modèles basés sur cette approche n'ont plus ce plafond de verre humain en terme de performances, mais peuvent aussi permettre de découvrir des relations insoupçonnées entre entrées et sorties, et ainsi créer de la connaissance. Cette translation de responsabilité de l'intelligence humaine vers l'intelligence machine a été le coeur des contributions de cette thèse. Nous proposons en effet quatre contributions afin de répondre à deux problématiques cliniques concrètes identifiées en introduction.

La première problématique clinique est, dans le cadre de la SCP, la prédiction des effets cliniques post-opératoires à partir de données préopératoires, afin de réaliser un outil d'aide à la décision pour la phase de sélection de patients. Le besoin clinique est réel : ce jour, les différents cliniciens intervenants sur la procédure (neurochirurgiens, neurologues, neuropsychologues...) se réunissent périodiquement afin de discuter des dossiers patients. Chacun utilise son expertise et ses connaissances afin de prédire les effets escomptés d'une éventuelle stimulation cérébrale profonde sur le patient. Après discussion, les cliniciens statuent de manière unanime. C'est une tâche difficile, mêlant objectivité et subjectivité, et dépendant de paramètres quantifiables, non quantifiables et souvent inconnus. Pourtant, le succès d'une stimulation cérébrale profonde dépend non seulement de la qualité de la chirurgie elle-même, mais aussi du sous-type et de l'état d'avancement de la maladie :

tous les patients ne répondent pas de la même manière à une SCP, rendant cette phase de sélection cruciale.

Dans l'objectif de proposer un outil d'aide à la décision clinique pour cette phase, nous avons proposé trois contributions :

— En premier lieu, il nous a paru naturel de nous intéresser aux tests et questionnaires cliniques réalisés par le patient avant l'opération. Les résultats de ces tests sont utilisés par les cliniciens afin de mieux comprendre la forme de la maladie du patient, son état d'avancement et ses particularités. L'information y est riche et cruciale, mais très difficile à exploiter par des algorithmes d'apprentissage machine. En effet, ces données sont nombreuses et, à cause de la réalité clinique du terrain, possèdent de nombreuses valeurs manquantes. Nous avons donc proposé une méthode basée sur les auto-encodeurs, une famille de réseaux de neurones artificiels profonds, afin de résoudre ces deux problèmes en même temps : le double objectif est que cette méthode soit capable de compresser ces données cliniques tout en étant robuste aux valeurs manquantes. Cette méthode a montré de bons résultats, surpassant deux méthodes linéaires sur les tâches de reconstruction des informations compressées et d'estimation des valeurs manquantes, et ce à différents degrés d'incomplétude des données d'entrée. Nous avons donc réalisé un système nous permettant d'exploiter plus facilement les informations cliniques issues des tests et questionnaires patient.

— En second lieu, nous nous sommes intéressés aux déformations du cerveau des patients, quantifiables grâce à l'imagerie par résonance magnétique. En effet, les déformations du cerveau sont très informatives quant à l'état d'avancement de la maladie du patient et sont utilisées par les cliniciens dans la prise de décision. Après étude de la littérature, nous avons constaté que les déformations de forme du striatum étaient représentatives de l'état de gravité de la maladie. Nous avons proposé une méthode se basant sur de l'analyse d'imagerie médicale, de la compression linéaire et de l'apprentissage machine afin d'étudier le potentiel et la pertinence de cette source d'information. Les résultats furent positifs. Cette méthode nous a permis de classifier, avec de bons taux de réussite, différentes cohortes de patients : des sujets sains, des sujets en phase prodromal (c'est à dire des patients en phase préliminaire de la maladie de Parkinson), des sujets diagnostiqués depuis moins de deux ans de la maladie de Parkinson ainsi que des sujets ayant réalisé une SCP, et donc à un stade plus avancé. Nous avons prouvé que le champ de déformations de surface du

striatum compressé était un biomarqueur précieux, et une bonne manière d'exploiter l'imagerie médicale.

— En utilisant les deux méthodes présentées précédemment, nous avons réalisé un système permettant de prédire les effets cliniques post-opératoires d'une SCP en fonction des informations préopératoires suivantes : les résultats des tests cliniques, les déformations du striatum ainsi que des données démographiques du patient. Un tel système est très novateur dans l'état de l'art : en effet, c'est le premier à considérer un nombre aussi important de variables d'entrée, mais aussi le premier à prédire une aussi grande variété de scores. En effet, nous avons évalué notre système sur 82 scores cliniques postopératoires (qu'ils soient moteurs ou neuropsychologiques). Nous avons prouvé que notre système était en mesure de prédire une majorité de ces scores, avec des coefficients de corrélations relativement haut pour certains, battant ainsi une méthode linéaire.

L'ensemble de ces trois contributions a permis de présenter des résultats novateurs et intéressants. Nous avons identifié plusieurs axes d'amélioration et avons réalisé une étude d'acceptabilité de notre système auprès de praticiens français, qui se sont montrés réceptifs et intéressés, nous encourageant à poursuivre nos travaux.

Nous nous sommes également intéressés à un second problème clinique survenant durant l'opération : la localisation du noyau cible par rapport à la trajectoire de l'électrode. En effet, les coordonnées de la cible définies avant l'opération ne sont pas assez fiables : le liquide cérébro-spinal fuit suite à la perforation de la boite crânienne, et la pression exercée par l'électrode sur le cerveau vient le décaler. De ce fait, l'équipe clinique utilise des enregistrements électrophysiologiques du cerveau par microélectrode. En analysant ces signaux auditivement, l'équipe clinique est en mesure de déterminer la localisation de l'électrode pendant l'enregistrement. En effet, les signaux électrophysiologiques émanant du noyau sous-thalamique, par exemple, ont des caractéristiques singulières reconnaissables à l'oreille par un clinicien entraîné. En répétant cette écoute à plusieurs profondeurs le long de la trajectoire, ils déterminent la localisation précise du noyau cible et peuvent ainsi continuer l'opération. Bien que très efficace, cette opération est aussi très longue, subjective et nécessite beaucoup d'expertise. En conséquence, l'automatisation de cette procédure a été un axe de recherche très fertile durant cette dernière décennie. Une étude de l'état de l'art nous a révélé que l'écrasante majorité des travaux à ce sujet sont basés sur des caractéristiques. Nous avons décidé d'innover en utilisant l'apprentissage profond, et plus particulièrement les réseaux de neurones convolutifs afin de classifier des

spectrogrammes, obtenus grâce à une transformée de Fourier à court terme, de signaux électrophysiologiques enregistrés par micro électrode. Nous avons obtenu de très bons résultats, dans le même ordre de grandeur que les résultats de l'état de l'art sur la localisation du noyau sous-thalamique en ne nécessitant qu'une seule seconde de signal. Notre structure de réseau de neurones faite sur mesure a démontré de meilleurs résultats que deux structures préexistantes de l'état de l'art en vision par ordinateur.

Dans un ultime chapitre, nous discutons des limitations gravitant autour de la recherche en apprentissage machine pour la stimulation cérébrale profonde et de son intégration dans la réalité clinique. Premièrement, nous notons que les plus gros enjeux ne se trouvent pas forcément du côté méthodologique, mais au niveau de la collecte des données. Ce jour, les bases de données cliniques sont de taille trop limité, trop rare et trop hétérogène pour garantir de bonnes performances, une comparabilité des méthodes et une bonne représentation de la variabilité des maladies traitées par la SCP. Nous notons également le besoin matériel et énergétique majeur induit par l'entrainement de réseaux de neurones profonds, ce qui représente une empreinte carbone non négligeable difficilement compatible avec les enjeux écologiques et énergétiques qui caractérisent notre époque. Enfin, nous notons que l'utilisation de systèmes d'aide à la décision clinique ne peut se faire qu'en satisfaisant plusieurs prérequis. L'outil proposé doit répondre à des critères de performance, de fiabilité, d'acceptabilité par les praticiens et les patients, soulève aussi des problèmes d'ordre éthique et dépend d'un cadre légal ce jour incomplet. Avant de conclure, nous évoquons les nombreuses perspectives techniques et cliniques en relation étroite avec les travaux que nous proposons.

PREAMBLE

This thesis takes place in the context of Deep Brain Stimulation (DBS) as a therapy for Parkinson's Disease (PD). DBS is an efficacious and promising way of treating abnormal movement disorders, such as PD, by electrically stimulating deep structures of the patient's brain, resulting in reduced motor symptoms and enhanced quality of life. However, the success of a DBS involves a lot of challenges, clinical problems and trade-offs, which depend on many variables. Many unknowns persist throughout clinical workflow, from screening to the procedure itself and the follow-up, creating an urgent need to develop computer assisted tools.

This thesis has been realized within the MediCIS team of the LTSI laboratory (UMR 1099 Inserm) in Rennes, France, between September 2017 and September 2020, under the supervision of Prof. JANNIN and Prof. HAEGELEN. The MediCIS team has been working on developing innovative computer assisted tools for DBS since 2013, with a focus on addressing concrete clinical problems and enhancing patient care.

To this extent, the team's recent works includes the development of a surgical planning software (PyDBS [1]), and methods to build anatomo-clinical atlases, i.e. maps showing the expected efficiency of stimulation locations around the anatomical targets of interest, based on several clinical criteria [2]–[6]. Specifically, the three works of Baumgarten *et al.* [4]–[6] used Machine Learning (ML) (a family of algorithms which automatically learn how to perform a task by retrospectively analyzing a database). This work improved the efficiency of planning DBS electrode placement and outperformed a traditional, non-ML-based state-of-the-art method. The MediCIS team's motivation has been to pursue data-driven approaches such as Deep Learning (DL) to develop new computer assisted tools for DBS, as the computational power and flexibility of DL have caused several breakthroughs in many research fields.

This thesis arose in this context and gave me the opportunity to work on important problems in a gripping and thriving working environment. During this thesis, I had the opportunity to propose four contributions, addressing two concrete clinical needs on the inclusion and surgery phases of DBS, allowing using machine intelligence to unveil knowledge from clinical data. This manuscript presents these contributions (Chapters 2, 3, 4, 5),

after and introductory chapter (Chapter 1) that explores DBS and ML, containing a systematic review of the use of ML in DBS and our hypotheses for this research. The final chapter (Chapter 6) evaluates these hypotheses and discusses some of the inherent limitations of ML in DBS as well as more long-term perspectives on the field.

List of publications

Published

International journals:

— **Maxime Peralta**, John S.H. Baxter, Ali R. Khan, Claire Haegelen, and Pierre Jannin. (2020). "Striatal shape alteration as a staging biomarker for Parkinson's Disease". NeuroImage: Clinical, 27.

International conferences:

— **Maxime Peralta**, Quoc Anh Bui, Antoine Ackaouy, Thibault Martin, Greydon Gilmore, Claire Haegelen, Paul Sauleau, John S.H. Baxter, and Pierre Jannin. (2020, July). "SepaConvNet for Localizing the Subthalamic Nucleus using One Second Micro-Electrode Recordings". In 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society in conjunction with the 43rd, Annual Conference of the Canadian Medical and Biological Engineering Society.

Submitted

International journals:

— **Maxime Peralta**, Claire Haegelen, Pierre Jannin, and John S.H. Baxter. (2019). "Data Imputation and Compression For Parkinson's Disease Clinical Questionnaires". Artificial Intelligence in Medicine (minor revision).

— **Maxime Peralta**, John S.H. Baxter, and Pierre Jannin. (2020). "Machine Learning in Deep Brain Stimulation". (to be submitted).

— Thibault Martin, **Maxime Peralta**, Greydon Gilmore, Paul Sauleau, Claire Haegelen, Pierre Jannin, and John S.H. Baxter. (2020). "Extending Convolutional Neural Networks for Localizing the Subthalamic Nucleus from Micro-Electrode Recordings in Parkinson's Disease". IEEE Transactions on Biomedical Engineering (submitted).

— Alicia Diot, **Maxime Peralta**, John S.H. Baxter, Pierre Jannin, and Claire Haegelen. (2020). "Exploring the acceptability of PassFlow, an AI based decision support software for Deep Brain Stimulation". Stereotactic and Functional Neurosurgery (submitted).

— John S.H. Baxter, **Maxime Peralta**, and Pierre Jannin. (2020). "Validating Medical Information Processing Algorithms in the Age of Machine Learning" (to be submitted).

— Elise Bannier, Giulio Gambarota, **Maxime Peralta**, Maud Guillen, Jean-Christophe Ferré, Tobias Kober, Anca Nica, Stephan Chabardes, and Claire Haegelen. (2020). "FLAWS imaging improves depiction of the thalamus subnuclei for DBS planning". Journal of Magnetic Resonance Imaging (submitted).

International conferences:

— **Maxime Peralta**, Claire Haegelen, Pierre Jannin, and John S.H. Baxter. (2020). "Multimodal Pre-Operative Biomarkers Can Predict Deep Brain Stimulation Outcomes". (to be submitted).

Chapter 1

Introduction

In this chapter, we will present in a first section Parkinson's Disease (PD), Deep Brain Stimulation (DBS) and its workflow in the Rennes University Hospital, and see how data-driven tools can be pertinent in this context. A second section will introduce Machine Learning (ML), specifically its interests and key concepts, and present how and why Deep Learning (DL) made data-driven methods trendy since 2012. Then, a systematic review of the literature for ML in DBS will be presented in a third section, in order to spot which recurrent limits could be overcome with an appropriate methodology. Finally, a last section will introduce our contributions and our main hypotheses.

1.1 Deep Brain Stimulation for Parkinson's Disease

1.1.1 Parkinson's Disease

PD is the second most common neurodegenerative disease, affecting around 1 percent of the population over 60 years old [7]. PD results from the degeneration of the dopamine-producing areas of the basal ganglia. Although classified as a movement disorder, it is known that symptoms extend far beyond motricity, including cognitive and neuropsychiatric symptoms [8], greatly impacting the autonomy and the quality of life of the patient.

There exists a large range of therapies for PD. It is important to note that none of them are technically a cure for PD, as they don't stop the progression of the disease but rather limit the symptomatology and enhance patients' quality of life. This disease is remarkably heterogeneous in its symptomatology and progression, and the response to treatments is unequal between different patients. In some cases, the disease is preceded by a prodromal phase, where some subtle symptoms (such as sleep disorder or hyposmia) can be detected for an early diagnosis (as illustrated in Figure 1.1).

Early in the course of the disease, pharmaceutical treatments are indicated. Unfortunately, past a certain stage of the disease, medication intakes can be too frequent or doses too massive, so continuous therapies can be used complementarily. While pump delivery of levodopa can be satisfying, its drawbacks make DBS sometimes preferable.

1.1.2 Deep Brain Stimulation

DBS is a common neurosurgical procedure, introduced in 1987 by Pr. Benabid [10], in which electrodes are positioned into deep regions of the brain to correct for abnormal neural behavior. Continuous stimulation of these regions typically greatly enhances the quality of life of the patient by limiting the motor symptoms. There exist three common bilateral targets for DBS for PD patients: the Ventral Intermediate nucleus of the thalamus (VIM), the Globus Pallidus internus (GPi) and the Subthalamic Nucleus (STN) [11]. Each of them is indicated for certain forms of the disease, and the expected clinical outcomes greatly differ according to the structure stimulated. For example, VIM stimulation is indicated only for trembling Parkinson's, whereas GPi and STN are considered for akinetic-hypertonic, or mixed forms of the disease. Where STN-DBS often leads to a

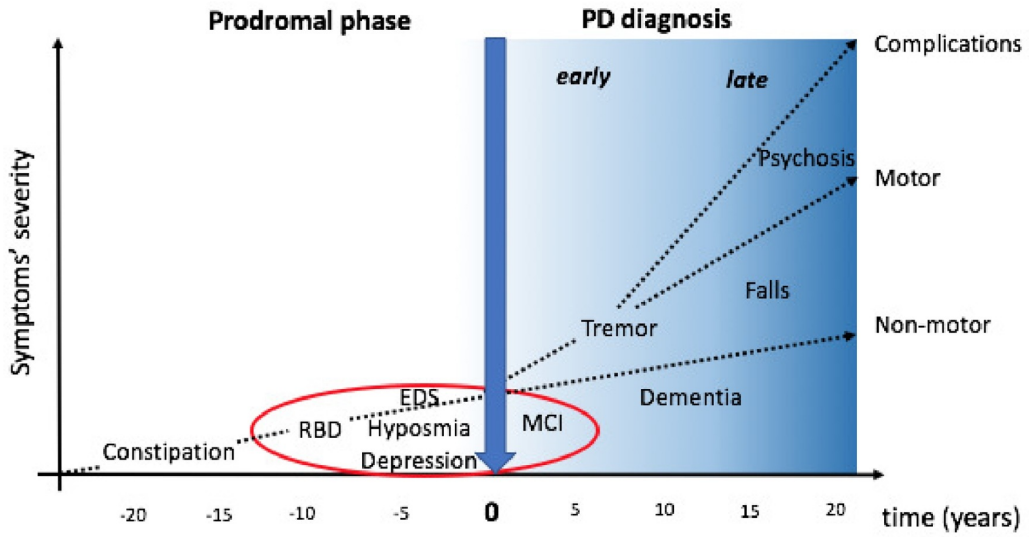


Figure 1.1 – PD is heterogeneous in its symptomatology and its progression, and can be preceded by a prodromal phase. Source: Amoroso *et al.* [9]

greater alleviation of motor symptoms, it also prompts undesirable side effects more [12]–[15].

DBS is an efficient but complex procedure, involving several steps, from patient inclusion to post-operative follow-up. The following section will present and detail DBS workflow in Rennes University Hospital's Department of Neurosurgery.

1.1.3 DBS Workflow in Rennes University Hospital

Figure 1.2 presents a synthetic workflow of DBS in Rennes, composed of four major phases. For each of them, the identified clinical problems for which computer assistance could be valuable are presented on the left. A problem written in blue means that this problem has already been addressed by the MediCIS team. A problem written in green means that it has been investigated in this thesis. On the right, data arising from these phases are noted. Each phase will be presented in the next paragraphs.

Inclusion

The inclusion is the first step of patient handling, as it consists in selecting patients who could benefit from undergoing DBS. This step is essential, as 1) there are more candidates for a DBS than surgical timeslots in the Rennes University Hospital, and 2) DBS is not indicated for all patients, because of the side effects the stimulation can

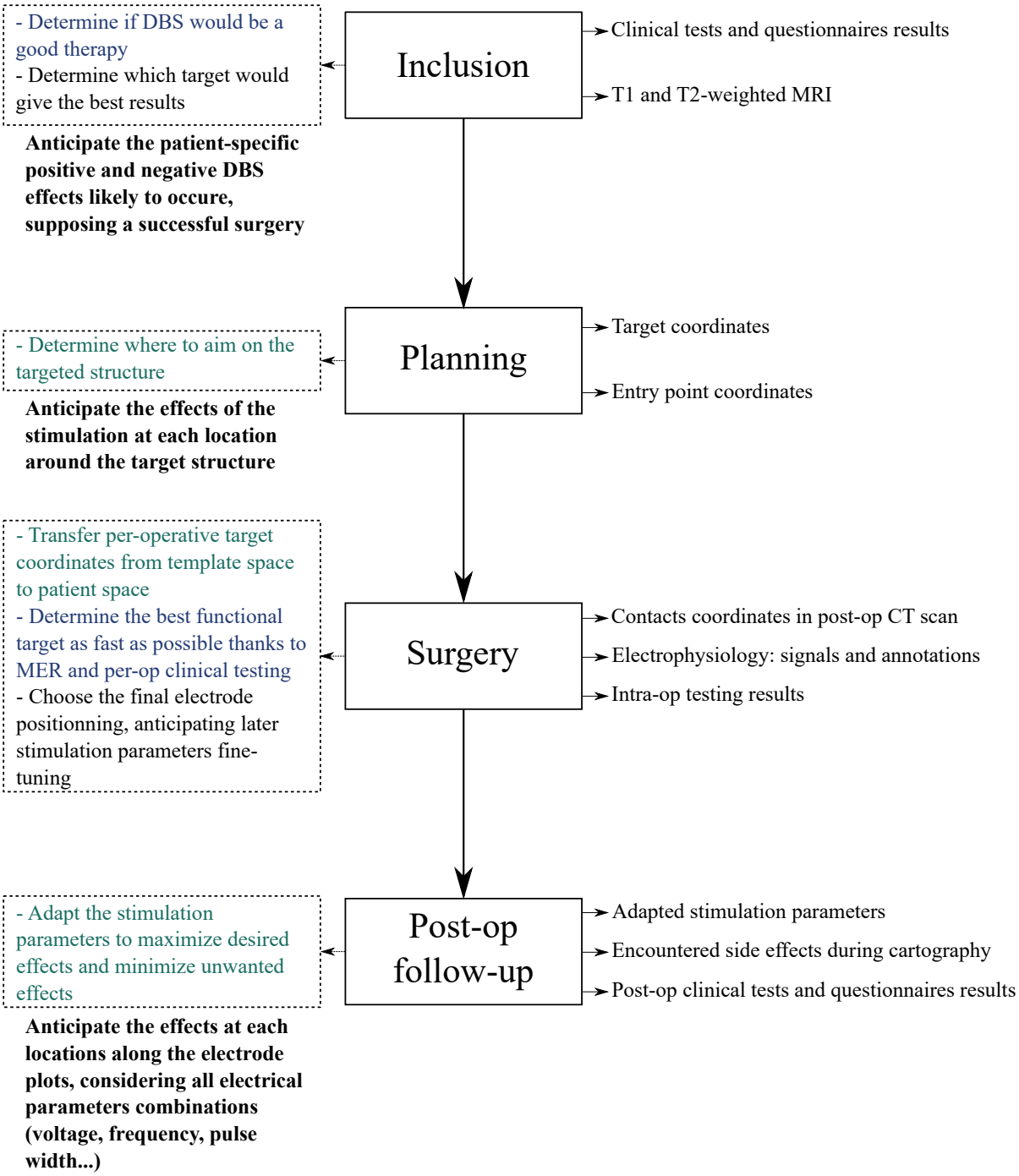


Figure 1.2 – Clinical workflow of DBS in Rennes University Hospital.

induce, and the risks caused by the surgery itself. To determine which patients should be included in the DBS protocol, clinician experts gather periodically in clinical *staff* and discuss the medical file of each patient. Several modalities of data compose this medical file, such as 1) clinical tests and patient questionnaires, 2) informal observations made by clinicians during patient's pre-inclusion visits, or 3) medical imaging such as T1-weighted Magnetic Resonance Imaging (MRI). In Rennes University Hospital, a DBS *staff* is typically constituted of up to ten experts (including neurosurgeons, neurologists, neuropsychologists, education nurses, residents of neurology...), each contributing with their domain of expertise. After discussion and debate, the clinical team makes a decision about the therapy of the patient, and determine which structure to target if DBS is chosen.

Clinical outcomes of DBS are a trade-off between positive effects and side effects and surgery risks. Therefore, the main clinical challenge during this phase is to be able to anticipate the effects of the stimulation on each anatomic target as accurately as possible, because any error could lead into taking a wrong, or suboptimal decision. Modeling the links between a patient's inclusion data and expected DBS results is a complicated task for clinicians, because this model is complex, largely unknown and relies on a lot of different modalities. As a simplification, clinician experts consider several per-modality known risk factors (such as important cortical atrophies, or poor response to levodopa), and sum them up to make the final decision.

This task is also challenging for the engineering community, but well suited to data-driven and ML analysis. In this thesis, we used information from clinical testing and questionnaires and imaging data to address this clinical problem. When it comes to clinical testing and questionnaires, the two main challenges are the high-dimensionality of the data, leading to a training problem known as the *curse of dimensionality* and missing values in clinical testing records, either randomly or *en bloc*, making data imputation almost mandatory. A multitude of anatomical and functional components lies in many imaging modalities, such as cortical atrophies, subcortical alterations or connectivity. Extracting these components necessitates powerful computer vision algorithms because the inter-patient variability is high and imaging scan are noisy, making the constitution of an informative, complete and consistent set of imaging biomarkers arduous.

Planning

Before the surgery, the neurosurgeon needs to carefully plan it. The two main objectives of this phase are to determine the surgery target position and the entry point for each

electrode (one or two). This can be achieved thanks to pre-operative imaging, typically T1 and T2-weighted MRI, and optionally anatomical or functional atlases if the structures of interest are not well observable on the patient images. The main clinical challenge is to determine the best stimulation target around, or within the targeted structure. It requires a fine knowledge of the possible side effects induced by the stimulation, such as the Pyramidal tract side effect (PTSE) [16]. The MediCIS team proposed several works in the past to assist the neurosurgeon on this challenge:

— Baumgarten *et al.* [5] proposed a predictive system to compute the threshold at which PTSE could occur.

— Baumgarten *et al.* [6] also proposed a predictive system to compute the therapeutic window, i.e. the voltage range which generates therapeutic effects and no side effects, at different stimulation locations.

— Lalys *et al.* [2] proposed a clustering method linking three motor and five neuropsychological scores with the spatial coordinates of the active contacts, leading to the construction of anatomo-clinical atlases for STN-DBS.

— Following this direction, Haegelen *et al.* [3] proposed new anatomo-clinical atlases for the STN and the GPi, and used a Volume of Tissue Activated (VTA) model to predict the effect of stimulation on four clinical scores at several locations around the target.

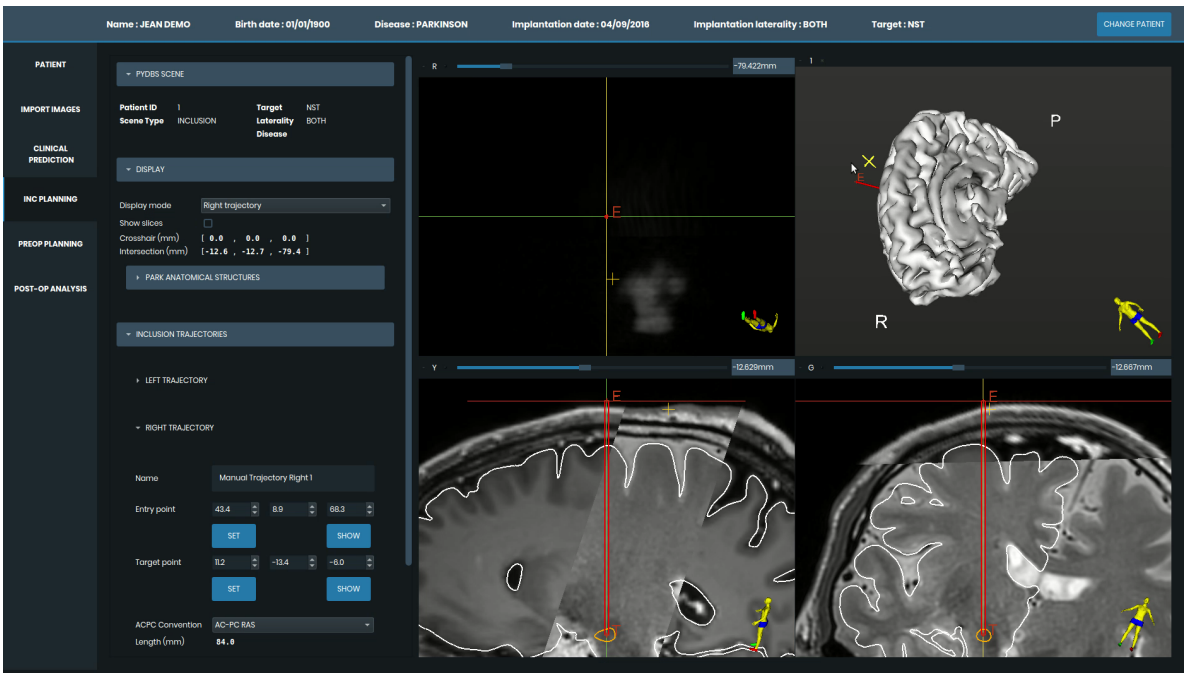
Some other technical issues can be encountered during the registration process, for example between T1 and T2-weighted MRI, or between the atlas and T1-weighted MRI.

In order to assist the neurosurgeon during the planning phase, the MediCIS team developed a software environment called PyDBS [1] (Figure 1.3) which addresses these technical points and integrates research works.

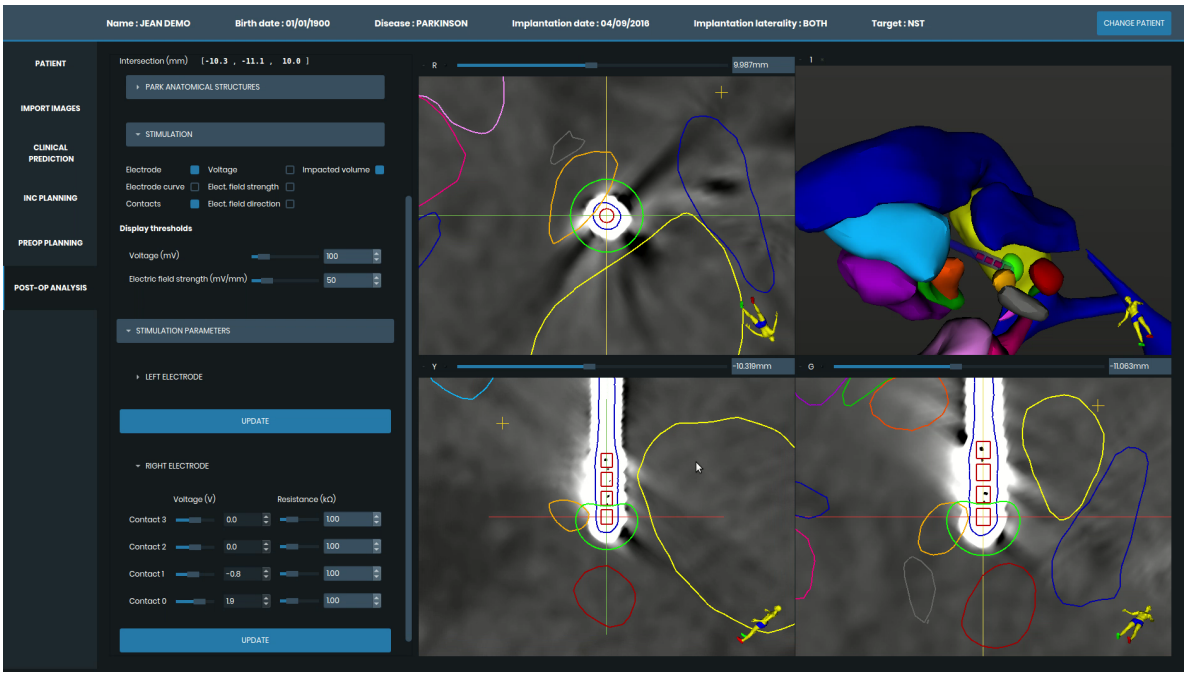
Surgery

The surgery itself is a complicated, multi-step operation involving a lot of decision-making, which won't be exhaustively described in this manuscript. The main phases are as follow:

First, a new round of medical imaging is acquired (typically Computerized Tomography (CT) scans) with a Leksell frame fixed on the patient skull (Figure 1.4). CT scans allow transferring the predetermined target coordinates and trajectory from the atlas



(a) Planning module of PyDBS.



(b) Post-operative module of PyDBS.

Figure 1.3 – Two modules of the PyDBS software developed by the MediCIS team. The first module (1.3a) is dedicated to assist the surgery planning. T2-MRI scan is registered on T1-MRI scan, as well as an anatomic atlas which shows the sub-cortical structures of interest (here, the STN in orange). From these views, the surgeon can define a target and a trajectory. The second, post-operative module (1.3b) automatically segments the electrodes on the post-operative CT-scan. The combination of a registered anatomical atlas and of a VTA model allows to visualize the effects of the stimulation on the anatomical structures.

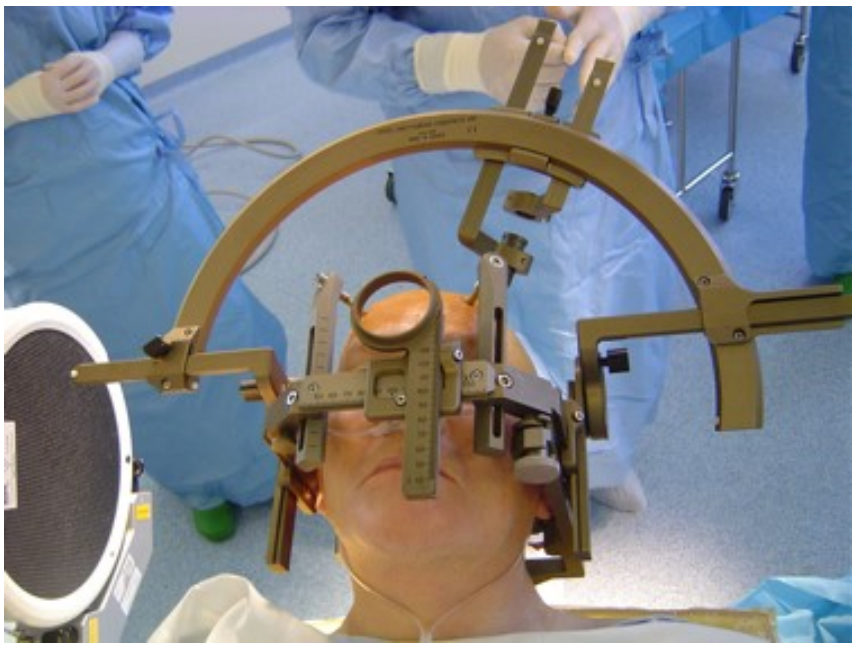


Figure 1.4 – The Leksell frame, which allows to accurately position the electrodes, is fixed on the patient's skull.

space to the patient CT space. On the same extent as during planning, registration issues can occur here.

After skull opening and the introduction of the microdrive, the position on the brain shifts because of the pressure and the loss of cerebrospinal fluid. In Rennes University Hospital, the position of the functional target is determined more finely thanks to Microelectrode Recordings (MER): by listening to the electrophysiological signals with a temporary electrode at several depths along three trajectories on the microdrive, the clinical team determines the STN borders regarding these trajectories. Even if effective, this step is time consuming considering the number of positions to test and the amount of time required at each position to qualify the nature of the signal. On a surgery that needs to be realized as fast as possible, both to limit risks and shorten the duration of discomfort of the patient, who is awake and not under medication intakes during the surgery, shortening this step can be greatly beneficial. Automatizing the recognition of the STN with MER has been extensively covered by the research community, and the results show that electrophysiological signals are complex and hardly reducible as a set of features.

Finally, once the position of the target has been refined, the stimulation electrode containing several stimulation contacts is introduced, and several electrical parameters are tested at several locations. The quality of each stimulation location is assessed by

Figure 1.5 – Several clinical tests are done during the operation, on the awake patient, to evaluate the quality of a stimulation spot.

considering the therapeutic window, i.e. the range between the voltage leading to positive clinical effect and the voltage inducing unwanted side effects (Figure 1.5). Once acceptable enough locations have been found, the final implantation of the electrode is chosen (Figure 1.6), with the care of giving some freedom regarding the centroid of the electrical stimulation, which can be important for adapting it during further post-operative fine-tuning.

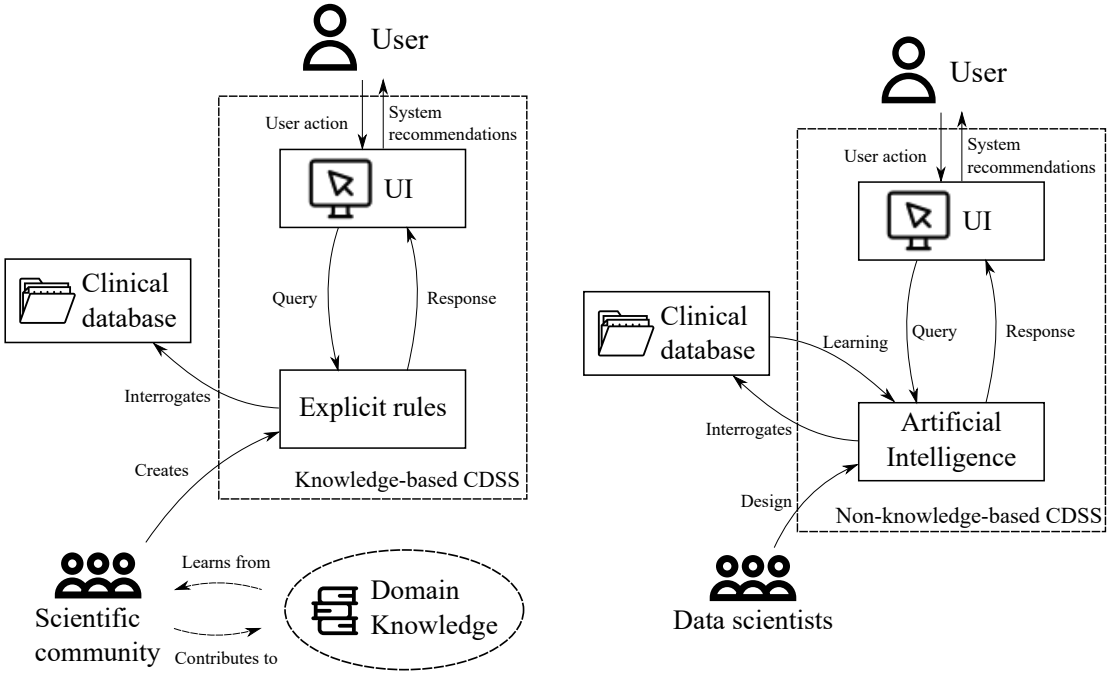
Post-operative follow-ups

After the surgery, the patient comes periodically to the hospital for a check-up. During these visits, cartography is done. Cartography consists in retesting different electrical parameters with different contacts looking for better clinical results. Indeed, the optimal stimulation parameters change with time for two reasons. Firstly, a phenomenon referred as lesional effect occurs: during the surgery, damages to the brain tissue around the target enhance the clinical symptomatology, which masks the effects of the stimulation itself. Up to three months after the surgery, once the tissue has recovered, the clinical state of the patient may have changed and new stimulation parameters must be found. Secondarily,

Figure 1.6 – Post-operative CT scan showing two electrodes stimulating the bilateral STN.

the stimulation not being a cure, the disease continues to progress leading to the necessity of increasing the voltage of the stimulation in order to keep beneficial effects few years later.

This cartography can be time consuming for the clinician in charge. Indeed, there are many parameters to tune (active contacts, voltage, frequency, pulse width...) and several clinical effects to check. This problem is amplified by the introduction of directional electrodes, which allows a more precise control of the VTA but typically multiplies the number of contacts by two [17]. From a data scientist point of view, this optimization problem could be addressed thanks to a data-driven method, but broad databases are complicated to gather. Moreover, the position of the electrodes has to undergo typically two registrations to be transferred into a template space (post-operative CT to patient T1 MRI, and T1 MRI to template space), which makes this crucial parameter very susceptible to error.



(a) Knowledge-based CDSS. (b) Non-knowledge-based CDSS.

Figure 1.7 – Comparison between knowledge-based CDSS and non-knowledge-based CDSS, according to [18]. In both cases, the user interacts with the user interface to query the CDSS core system. The system then interrogates the clinical database to retrieve the appropriate patient clinical information, and returns a response to the user interface. In the case of knowledge-based CDSS, the core system is a set of rules explicitly created by experts. In the case of non-knowledge-based CDSS, the core system is based on artificial intelligence and machine learning and learns the knowledge from the retrospective clinical database.

The concept of Clinical Decision Support Systems

Due to the highly complex and multi-modal nature of the procedure, each step of DBS presents challenges to both the human expert and to machine learning. A lot of expertise and domain knowledge is crucial for the success of this procedure. Since the 1980s, computer-assisted tools, referred to as Clinical Decision Support Systems (CDSS)s, have been designed to support clinicians in neurosurgery [18]. CDSSs are commonly divided in two categories: knowledge-based and non-knowledge-based. In knowledge-based CDSS, all of the tool's intelligence arises from the human. Decision rules are explicitly programmed according to the medical knowledge [19], and the purpose of the CDSS is solely to retrieve

the data, to evaluate the rule and to display the result with a User Interface (UI). In non-knowledge-based CDSSs, machine learning replaces expert medical knowledge in order to address new challenges and to reach new levels of performance. Figure 1.7 compares the two approaches. In knowledge-based CDSSs, we can see that the CDSS's intelligence is encoded as a set of rules explicitly created by the clinician experts thanks to their interaction with the domain knowledge. For the non-knowledge-based CDSSs, the intelligence is artificially generated by learning from a database using ML tools designed by data scientists, without requiring external expert knowledge.

Pioneering works in computer-assisted DBS

In the early 2000s, the first works towards assisting the planning phase of DBS with CDSSs appeared, relying on anatomical or functional atlases. The main motivation was that, at that time, the low spatial resolution of medical imaging scans prevented a direct location of the structures of interest, leading to an extensive intra-operative electrophysiological exploration. Such atlases, once registered to the patient's scan, allowed for an indirect location of the anatomical structure. Castro *et al.* [20] compared experts against atlas-based methods in terms of the location accuracy of the STN, and benchmarked several non-rigid registration algorithms. A first direction was the creation of anatomical atlas thanks to histological data [21], or histological data fused to MRI scans [22], [23]. Other works proposed to enrich planning systems with other data modalities. The fusion of several anatomical and functional data modalities registered to the individual patient images can enhance the accuracy and the generalization ability of such systems. Guo *et al.* [24] proposed a visualization and navigation system which simultaneously shows a segmented atlas of the deep brain nuclei, an electrophysiology database and a collection of post-operative surgical targets of previous patients. D'Haese *et al.* [25] proposed a similar system, including an intra-operative interface which allows to use it prior, but also during the operation, with the ambition of reducing the expertise required to perform DBS and therefore increase the number of procedures per year.

Section conclusion

Deep brain stimulation is a complex, multi-modal procedure with multiple steps. Each of these steps offers clinical and engineering challenges, often interlinked. The stakes of the procedure motivate the development of CDSS to enhance its safety, reliability and success. The multi-modal nature of these challenges makes data-driven approaches relevant and promising. However, the reliance of engineering research on clinical prior is likely to limit the scope of potential applications. In the next section, we will introduce the concept of machine learning, and explain the new trend of data-driven methods led by the rise of deep learning.

1.2 Key concepts of Machine Learning

In this section, we will introduce the concept of ML, its interesting characteristics and limitations. We will then focus on bottom-up, pure data-driven approaches before presenting DL and how it became trendy, and conclude regarding the purposes of such approaches in medical research.

1.2.1 Objectives of Machine Learning

The term 'Machine Learning' appeared for the first time in 1959 in the works of Samuel *et al.* [26]. It is a branch of Artificial Intelligence (AI) (Figure 1.8) which consists in letting an algorithm learn to perform a task thanks to a database of experiences without requiring any explicit programming of the user. On its most common form, called supervised learning, it consists in predicting an output variable (either categorical, for *classification* tasks, or continuous, for *regression* tasks) from a set of inputs, called features. To model the link between the inputs and the output, the algorithm processes a dataset of multiple known pair of inputs and outputs. If the training has been successful, the trained model is expected to be able to predict the unknown output of a new sample from its known inputs.

The other common form of machine learning is called unsupervised learning, where there is no output variable to predict. In this case, the model performs a task solely on the set of inputs, such as clustering.

There are two main purposes in using a ML model:

— The first one, similarly to classic statistics, is to measure how much a set of explanatory variables is correlated to a target variable. In this case, the interest of ML is its ability to take advantage of large databases and its higher ability to model non-linear relations between variables compared to classic statistics [27]. The ability of the ML model to predict unknown samples successfully, by comparing its predictions to the ground truths, represents the degree of correlation between inputs and output.

— The second, most common purpose is toward using the model prospectively. The model will not be restricted to a retrospective analysis of a database, but rather execute the prediction task prospectively providing that the test performance of the model is satisfying enough, allowing for the automation of the process.

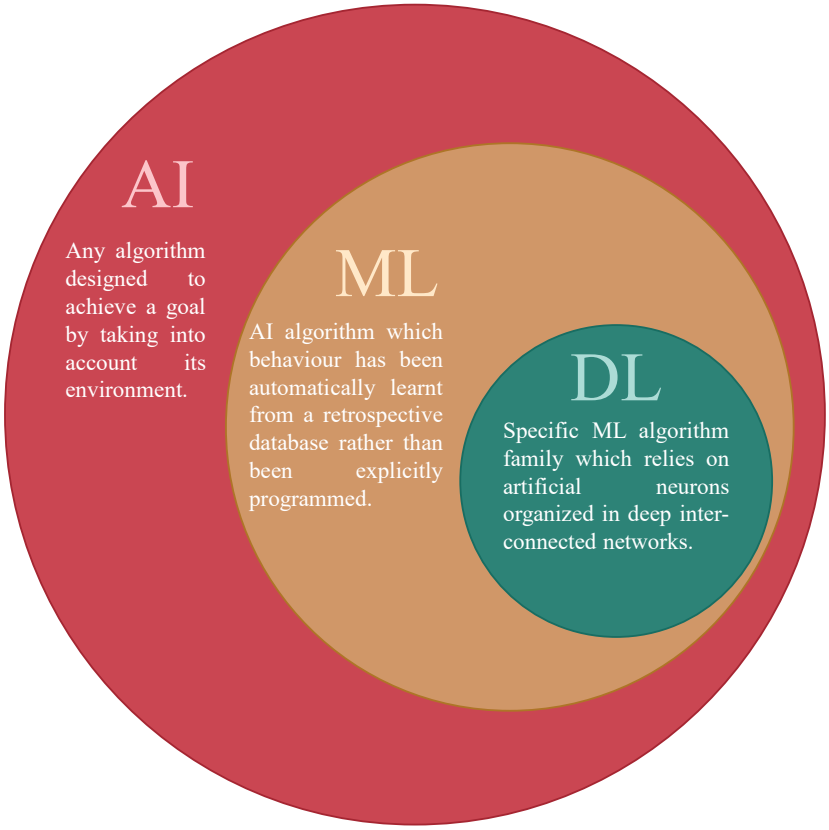


Figure 1.8 – Deep learning (DL) is a subset of Machine Learning (ML) algorithms, which itself is a particular way of doing Artificial Intelligence (AI).

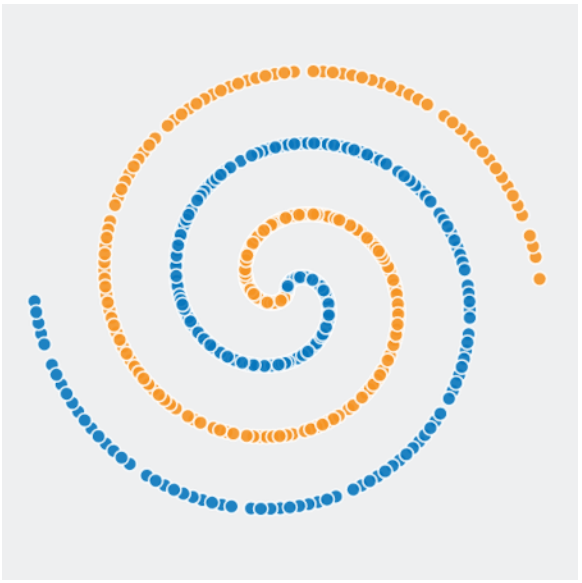
In both cases, a validation method is required, necessitating at least two datasets. The first one is the training dataset on which the model will learn, and the second is the testing dataset, on which the model's predictions will be compared against the ground truth, allowing the have a metric representative of the performances of the model on unseen data. It is also recommended to use a third dataset, called validation dataset, during the training process in order to monitor it or to optimize the model's hyper-parameters. This data separation is crucial as ML models are usually able to learn *by heart* the training database, making close to perfect predictions for every sample. Nonetheless, it does not signify that the model effectively *understood* the data and would be able to generalize its learning for new data samples.

On real situations, this data splitting can be problematic because it requires training, validation and testing sets to be representative of the data distribution. In medical research, where databases are likely to be limited in size, the most common evaluation strategy employed is the k-fold Cross-Validation (CV). K-fold CV consists in randomly splitting the data in k folds, taking $k-1$ of them as the training set, and the remaining one as the testing set. The model is trained on the training set, evaluated on the testing set, and this entire process is repeated k times until every fold has been on the testing set once. The performance on all the k testing fold are then pooled together, allowing for the obtained metrics to be representative of the performance on the whole database while ensuring that the model has been evaluated on unseen data only. High values of k give better estimates, as more data are available for training, but take longer to compute.

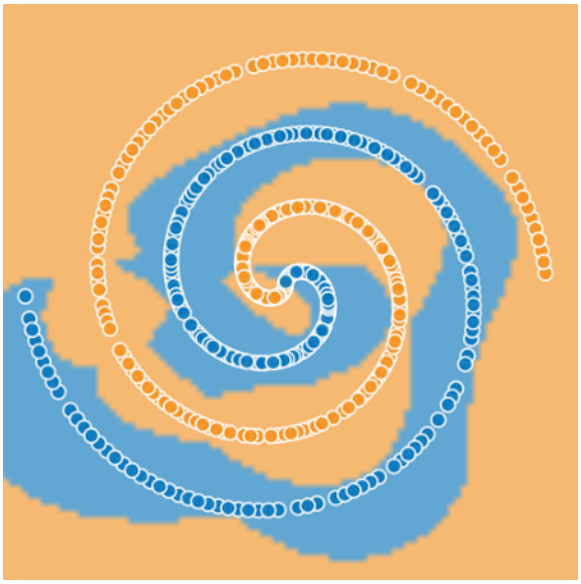
1.2.2 Limitations

The major condition for a ML model to be accurate is the amount of training data available. The more complex the input distribution is, the more input data will be necessary for the model to 'understand' it.

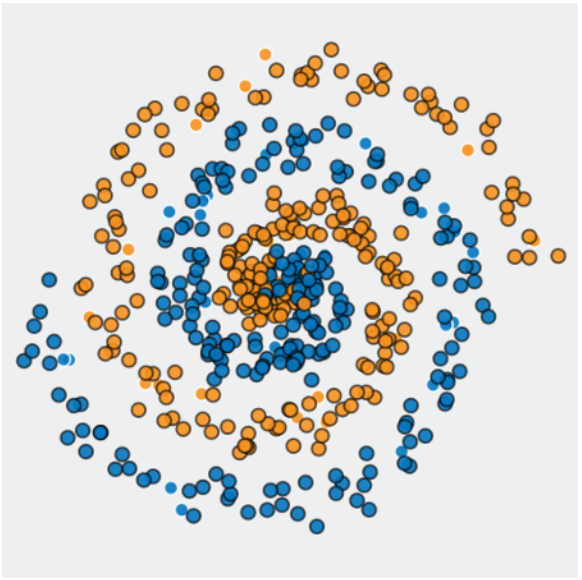
To illustrate this constraint, we conduct a simple experiment using an online tool proposed by Google, one of the main firms in the ML and DL research. This tool, called *TensorFlow's Playground* (`https://playground.tensorflow.org`) allows visualizing the training of simple Artificial Neural Network (ANN)s on various toy examples. Here, we train a small ANN to 'understand' the distribution of a 2D spiral, based on some simple features computed from the 2D spatial coordinates, through a classification task. Figure 1.9a shows the training data distribution. The spiral is decomposed in two classes (orange and blue), and the ANN has to learn a decomposition of the 2D space to discrim-



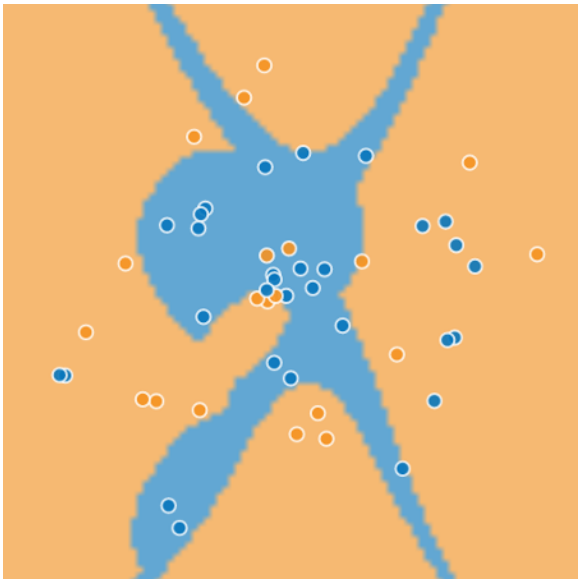
(a) Original, clean distribution of the input data.



(b) Segmentation of the clean input space by the Artificial Neural Network after training.



(c) Noisy distribution of the input data.



(d) Segmentation of the noisy input space by the Artificial Neural Network after training on 10% of the data.

Figure 1.9 – TensorFlow: playground experiment, showing how a simple Artificial Neural Network learns how to segment a 2D space, between 'blue' and 'orange' zones. 'Blue' and 'orange' training points are drawn from the distribution of a 2D spiral. Input features are simple functions of the spacial coordinates of the points.

inate orange from blue dots. Figure 1.9b shows how the ANN segmented the 2D space after training, the blue (resp. orange) background showing which location of the 2D space the ANN considers as belonging to the 'blue' (resp. 'orange') class. We can see that no points are in the opposite color space; therefore, there is no misclassification: the ANN understood the input distribution perfectly. Unfortunately, such a classification case is not realistic: in real world problems, the input data is usually noisy (because acquisition systems are unlikely to be perfect) and there is a limited amount of training data points. Both problems can be simulated with TensorFlow's Playground. Figure 1.9c shows the noisy distribution of the same spiral. Additionally, this time, the ANN have to learn solely from 10% of the dataset. The result, after training, is shown in Figure 1.9d. We can see that, this time, the ANN completely failed to catch the distribution of the input data. This result is expectable as even a human couldn't guess the original spiral distribution from the few training data points remaining. This is a problem that a much wider, more powerful ANN couldn't solve: the problem is due to the training data distribution which is too sparse in the input space.

This sparsity problem is even more problematic in high-dimensional feature spaces, because of a phenomenon called the *curse of dimensionality*.

For these reasons, the most common strategies employed by data scientists were either to isolate and consider a few features at a time, or to project the input data into a smaller, simpler feature space, making the task of the ML model easier. This projection is commonly done by handcrafting appropriate features for the data and the problem.

1.2.3 From feature engineering to data-driven approaches

We can qualify an approach of data-driven, or bottom-up by its propensity of letting the knowledge arises directly from the data thanks to powerful ML algorithms. It is opposed to top-down approaches, where the data is used solely to validate pre-existing hypothesis originating from the human expert.

We can reasonably say that the level of complexity the human brain can consider is fundamentally capped. Passed a certain amount of information to consider simultaneously, a human can hardly find inter-correlation in variables to generalize a phenomenon.

As stated in the previous section, a common strategy for data scientists to address a ML problem is to intellectualize it and propose efficient ways to synthesize the raw information into understandable features, making the input space less complex and smaller in dimensionality. A feature is a direct transcription of what a domain expert or a com-

munity can understand of data, and of its ability to be relevant relative to the task at hand. In this extent, feature-engineering is highly depending on domain knowledge and expertise. Therefore, past a certain level of maturity in a research field, the ability of the community to find more and more representative and accurate set of features will stagnate to a level where it is hard to find any new feature allowing to have a drastic effect on performances. The major limit in term of performance would therefore be human expertise rather than the power of ML algorithms themselves. In this context, employing a data-driven paradigm, by taking the information closer and closer to its raw form, i.e., not synthesizing the input data but feeding it as straightforwardly as possible unlocks potential new levels of performance, translating the responsibility to recognize low-level patterns in the data to the machine, which, by essence, can scale this ability to higher levels of dimensionality and complexity.

The works of Baumgarten *et al.* [5] offer a good example the interests of such a paradigm. Baumgarten *et al.* compared two approaches to predict the occurrence of PTSE for DBS, an undesirable clinical side effect of electrical stimulation of the STN:

— A feature-based approach, which consists in computing an isotropic VTA and measuring its intersection with the internal capsule. By doing so, the classifier works on a simple, unidimensional input space, which is likely to be straightforwardly correlated with the clinical outcome, as it is known that PTSE is happening when the internal capsule is affected by the stimulation. Nonetheless, the anisotropic property of the neuronal tissue surrounding the internal capsule is not covered with this approach, because the feature used to synthesize the raw information does not take it into account.

— A pure data-driven approach where the classifier learns directly from the spatial coordinates of the electrode as well as the voltage of the stimulation. While the input space has here four dimensions and is more complexly linked to PTSE occurrence, the classifier has access to the complete, raw source of information and is, in this extent, able to learn how to transform this raw information into a more meaningful set of features. With this approach, the classifier could theoretically implicitly *learn* the anisotropy of the tissue surrounding the internal capsule (in other terms, deduce it directly from the database without requiring it to be explicitly measured), and take advantage of this property to be more accurate in its prediction than the first system.

Baumgarten *et al.* [5] showed that the second approach outperformed the first one, proving that it could do a better utilization of the raw input information than the feature-oriented one. In order to improve the feature-oriented method, it would be necessary to go through a new complex, time-consuming feature engineering step by, for example, computing the anisotropy of the tissue and update the VTA model accordingly. Even by doing so, it is not warranted that this update would ensure superior performance from the feature-oriented approach, as the reason for data driven method's better performance could in fact be due to yet another implicitly learned property than anisotropy.

Even if the translation of complexity from the human to the machine can be beneficial, it also comes at a cost. The more the input space is high dimensional, raw and complex, the more data and computational power is required by the machine, generally more than a human would need. A good example for this phenomenon is Reinforcement Learning (RL) (a sub-field of AI used to train an agent to interact with an environment, sometimes relying on DL). A well-known breakthrough of the AI community is the AlphaStar agent [28], an algorithm developed by Deepmind, using DL and RL which was able to beat professional players of Starcraft 2, a real time complex strategy video game. Before AlphaStar, this task was considered as impossible for classic AI algorithms because of the tremendous amount of real-time information to process and analyze. The particularity of AlphaStar is its ability to learn from experiences by playing against itself, without requiring a human to teach it the game's rules or human strategies, nor hard-coding its behavior. It is this translation of game intelligence, from the human developers to the algorithm itself, that makes AlphaStar a good example of how bottom-up, data-driven approach can lead to new levels of algorithm performance, in this case even topping expert human level. Nonetheless, to achieve human-comparable levels of performances, AlphaStar had to train during the equivalent of **200 years** of real-time StarCraft play, while a human can reasonably achieve this in a few thousands of play time: to outperform an expert, AlphaStar needed to learn from about a thousand times more experiences than him.

1.2.4 The rise of Deep Learning

The increased interest in bottom-up, non-feature engineering-based approaches is usually attributed to computer vision, a research field revolutionized by DL in 2012. DL is a family of ML algorithms (Figure 1.8) based on composing layers of simple computational units, called *neurons*. Organizing these neurons in broad and deep networks allow

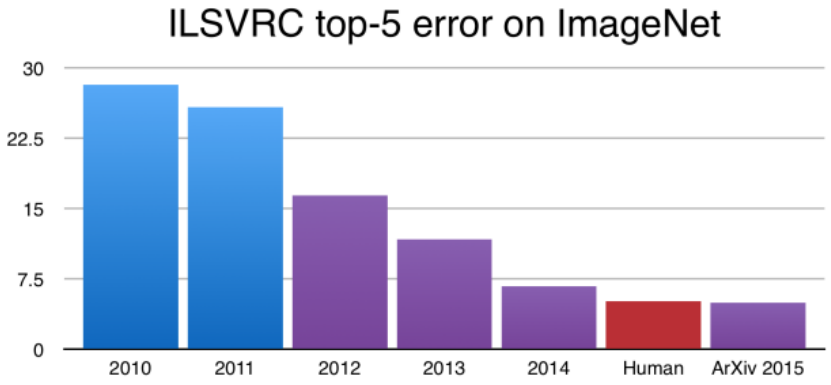


Figure 1.10 – Top-5 error rate of the best algorithms at ImageNet recognition task for the ILSVRC, yearly between 2010 and 2015. In blue: feature-based methods. In purple: DL-based method. In red: human error rate. Source: `https://devblogs.nvidia.com/mocha-jl-deep-learning-julia/`.

computing highly non-linear relations between the input and output, showing greater computational power than other traditional ML models.

ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [29] is an annual contest consisting in various CV tasks, such as object recognition or segmentation on the ImageNet dataset. Figure 1.10 presents the top-5 error rate of the best algorithms every year between 2010 and 2015.

Up to 2011, all the approaches were feature-based, achieving at best error rates of 0.2577 in 2011. The best performances were obtained by using a ML method learning from Scale-Invariant Feature Transform (SIFT) points extracted from the images [30], which reduces the high-dimension images to a set of a few points' coordinates. A huge breakthrough happened the next year, as the SuperVision team of the University of Toronto used for the first time on this challenge a deep Convolutional Neural Network (CNN) learning from the raw images directly. After this year, more and more CNN were proposed and the results got better and better until beating human-level error rate in 2015.

Similarly, DL achieved various breakthrough in other key research fields such as Natural Language Processing (NLP) [31], or artificial data generation [32] using other ANN topologies.

It is of common knowledge that DL's performance better scales with dataset size than other ML algorithms. Moreover, the ability of DL to construct high-level features from the raw data (a phenomenon known as *Representation Learning*) makes it a first choice tool when it comes to data-driven approaches. Purushotham *et al.* [33] remarkably

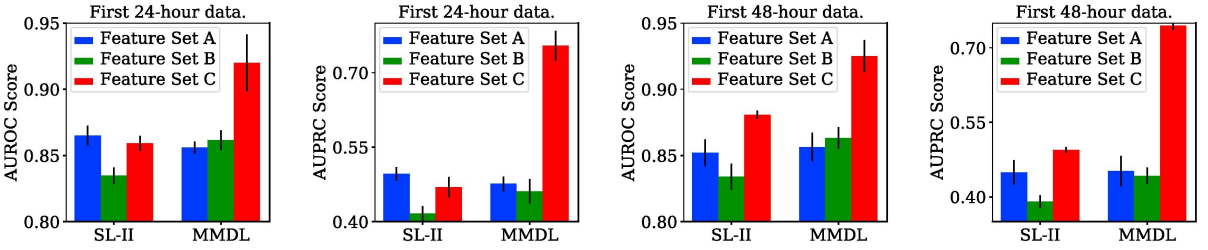


Figure 1.11 – Deep Learning ('MMDL') performs better than machine learning ('SL-II') on the feature set C, which corresponds to the rawest form of the data, in opposition of feature set A, which is composed of high-level, hand-crafted features. Source: [33].

quantified both effects by comparing the performances of a DL method to a ML method with several features set, on several clinical prediction tasks and with large datasets. Figure 1.11 presents the results of an experiment carried on this paper. A multimodal deep learning method ('MMDL') is compared with a powerful 'SuperLearner' (SL-II) which combines the predictions of several machine learning algorithms on an in-hospital mortality prediction task, with three feature sets:

— Feature set A is composed of 17 hand-designed high-level features (which is a feature engineering-based approach).

— Feature set B is composed of 20 raw and unprocessed selected features directly related to the 17 features of set A (which is a data-driven approach).

— Feature set C is composed of a large pool of 136 raw features (which is an even more extreme data-driven approach).

We can see that, while the performance of DL and ML are similar for hand-crafted features, they are drastically better when a raw and large pool of features is used in conjunction with a DL model.

DL (and more generally the whole ML field), began to raise the interest of the public, research founders and communities, as well as journals and conferences, as shown in the 2019 AI Index Report of Stanford University [34]. This report notably mentions that:

— "The share of published AI papers in total papers has grown three-fold in 20 years, accounting for 3% of peer reviewed journal publications and 9% of published conference papers" [34]

— Major artificial intelligence conferences have seen their attendance being multiplied by up to eight times between 2012 and 2019, as shown in Figure 1.13.

Figure 1.12 – Google trends of the term 'Deep Learning' between January 2004 and May 2020.

— "The share of earning calls where AI is mentioned has increased substantially, from 0.01% of total earnings calls in 2010 to 0.42% in 2018" [34]

In France, we can also mention the report [35] of the deputy Villani leading the government to intend to invest 1.5 billion euros in AI (including 700 million euros for public research) by the end of 2022. Lastly, the frequency of 'Deep Learning' as a Google search query can be quantified thanks to Google Trends, and is a good indicator of the general public interest and of the trendiness of the concept. Figure 1.12 shows the number of 'Deep Learning' search queries between 2004 and 2020.

DL influence also extended to healthcare research. A well-known breakthrough of DL in healthcare research was the CNN proposed by Esteva *et al.* [36] which outperformed the performances of some human experts for skin cancer detection. Recently, Miotto *et al.* [37] reviewed 32 DL papers in healthcare, and pointed out four families of DL structures (CNN, Recurrent Neural Networks (RNN), Restricted Boltzmann Machine (RBM) and Auto-Encoder (AE)) that encountered success in healthcare, and notably stated in conclusion as a key point that "early applications of deep learning to biomedical data showed effective opportunities to model, represent and learn from [...] complex and heterogeneous sources."

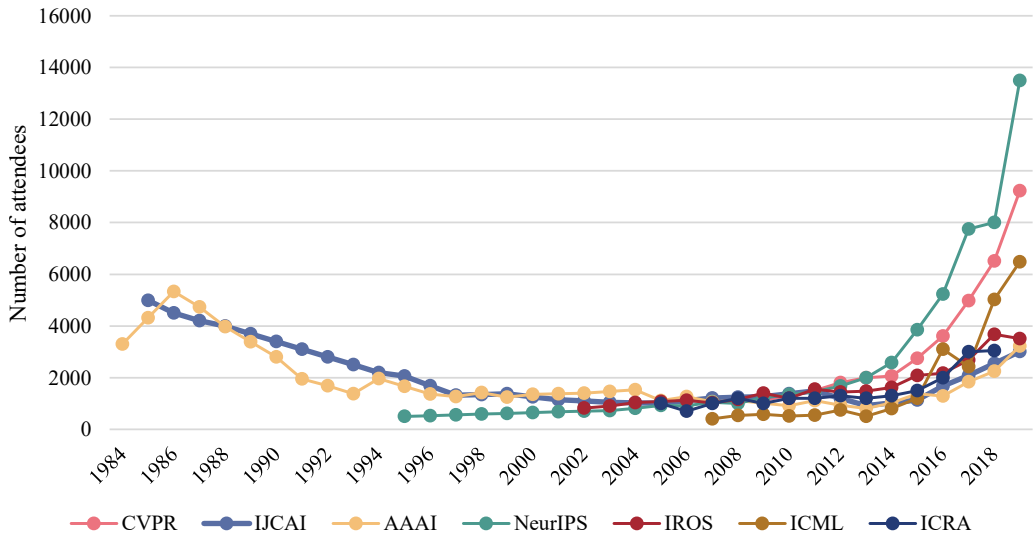


Figure 1.13 – Number of attendees of some major artificial intelligence conferences between 1984 and 2019 (data source: [34]).

Section conclusion

Machine learning is a mature research field which consists for an algorithm to automatically learn to perform a task by processing a retrospective database. Once the model has learned, it is expected to be able to generalize its knowledge to new, unseen samples prospectively. Its ability to draw information from complex, multimodal data and to perform complicated tasks at, or above human performances make ML an important component of non-knowledge-based CDSSs. Machine learning's first limitation is that, for complex and high-dimensional problems, it requires a large number of data samples to ensure a good ability for the model to *understand* the input distributions. Data scientists made great efforts to reduce both the complexity of the input distribution learned by the model and the dimensionality of the input space by handcrafting and selecting features tailored for every problem, or by isolating a few explanatory variables at a time. Nonetheless, passed a certain point of maturity, this strategy fundamentally caps the performances of the model because the limitations arise from feature engineering itself and therefore human intelligence rather than from the ML model, motivating us to develop data-driven approaches instead. On the next section, we will see how data-driven approaches have been used on deep brain stimulation so far.

1.3 Machine learning in DBS: a systematic review

> "In the last few decades, the volume and complexity of bio-medical data have grown beyond the physician's ability to extract all meaningful data patterns using conventional statistical methods alone. [...] The complex diagnostic and therapeutic modalities used in neurosurgery provide a vast amount of data that is ideally suited for ML models." Senders *et al.* [38]

The interest of ML to assist clinicians in healthcare has been underlined for a long time, and several reviews have been conducted on the matter.

Celtikci *et al.* [27] and Bulchlak *et al.* [39] conducted systematic reviews of ML to assist the decision-making in neurosurgery. Senders *et al.* also realized a review of ML in neurosurgery [38], and proposed another one to compare the performances of ML to human experts for diagnosis, surgery planning or outcome prediction tasks in neurosurgery [40].

These reviews highlighted interesting ML applications, but none of them focused on DBS specifically. Utilization of ML in DBS is a broad area of research as the methods, data modalities and tasks are numerous. In the prospect of drawing a synthetic landscape of this research field, extracting major trends, better identifying recurrent methodological limits, and thus better locate works presented in this thesis, we conducted a systematic review focused on ML in DBS. The next sections will present the methodology employed to identify the corpus of paper to analyse, the data acquired on each paper, the results of the analysis and a discussion about these results.

1.3.1 Material and Methods

Figure 1.14 presents the workflow used to select papers for further analysis.

Paper search strategy

We search relevant papers in the literature by making three queries on two search engine, on June 30th 2020:

— PubMed, with MeSH method, with the following query: ('Machine Learning'[Mesh] OR 'Artificial Intelligence'[MeSH:noexp] OR 'Neural Networks, Computer'[MeSH]) AND 'Deep Brain Stimulation'[Mesh]

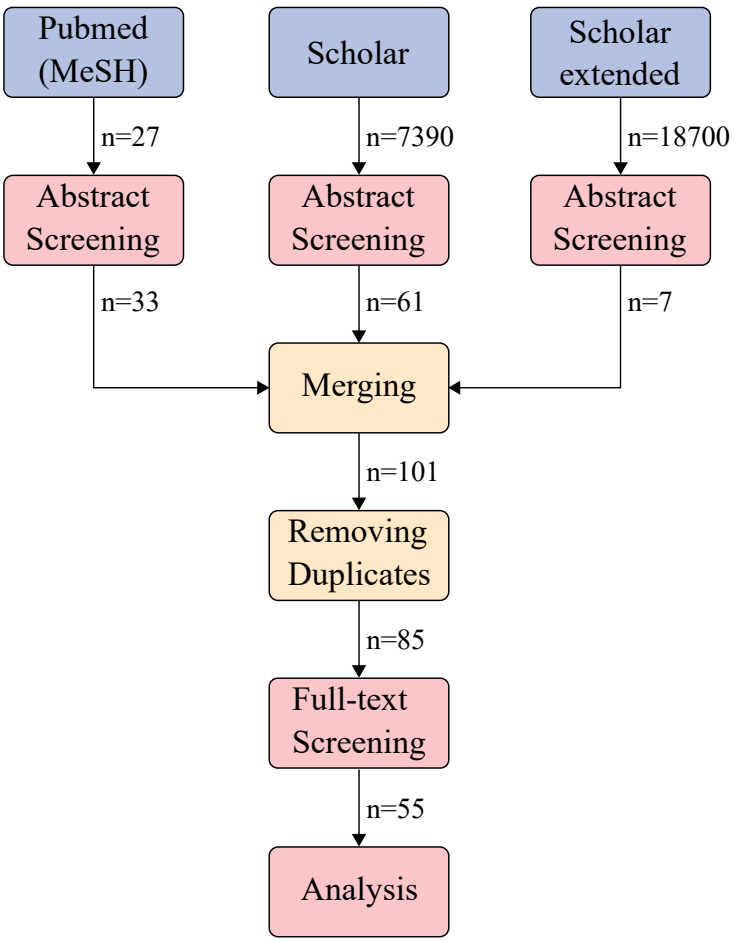


Figure 1.14 – Workflow to select the corpus of 55 papers to classify.

— Google Scholar, with the following query: (‘machine learning’ OR ‘deep learning’ OR ‘data-driven’ OR ‘learning-based’ OR ‘artificial intelligence’)(‘deep brain stimulation’)

— Google Scholar, with the following query: (‘prediction’)(‘deep brain stimulation’)

We chose to use PubMed with MeSH method as it is a proven way of browsing papers in medical research, providing that they are tagged with the appropriate MeSH terms. We did not add the MeSH term ‘Deep Learning’ to the query as it is already included in the ‘Machine Learning’ hierarchy. We turned off automatic explosion of the MeSH term ‘Artificial Intelligence’ in order not to include other unrelated topics such as ‘Robotics’. We made two different queries on Google Scholar as it is more comprehensive than PubMed. The first one was composed of targeted keywords, and the second one consisted of a broader term (‘prediction’) in order to include additional papers that could have been

missed by the first two queries. Due to the high number of papers returned by queries on Google Scholar, we analyzed the results page by page and stopped when we retained no new papers two pages in a row after title and abstract screening (the results being sorted by relevance). We decided not to merge both Google Scholar queries into a single as the term 'prediction' is broad and returned a lot of irrelevant papers. Therefore, including this term in the first Google Scholar query could have made relevant items sparser.

In order to keep a fully systematic methodology, to avoid flaws in the results and to make our screening method reproducible, we chose not to manually include additional papers in the corpus.

Selection process

The first author screened each paper by reading the title and abstract, with the following criteria:

— The paper has to be methodological: it must at least validate a method.

— ML has to be at the heart of the methodology employed.

— The paper has to be clinically validated on patients. Therefore, papers validated with synthetic data, or using a non-human cohort were discarded.

— The paper has to address a clear and identified clinical problem.

— The paper has to be peer-reviewed. If we couldn't obtain the published version, the pre-print version was used. Thesis manuscripts and reviews were discarded.

— The paper has to be written in English.

The number of papers returned by each query and the number of papers kept after title and abstract screening is indicated in Figure 1.14. We merged the results provided by the three queries, removed duplicates and obtained a corpus of 85 papers. The first author then re-screened each paper by reading the full text according to the same selection criteria. Thirty papers were discarded (notably six papers for being a preliminary version of another paper retained).

Data obtained from each paper

Each of the 55 paper composing the final corpus was described according to four classes, as presented on Figure 1.15.

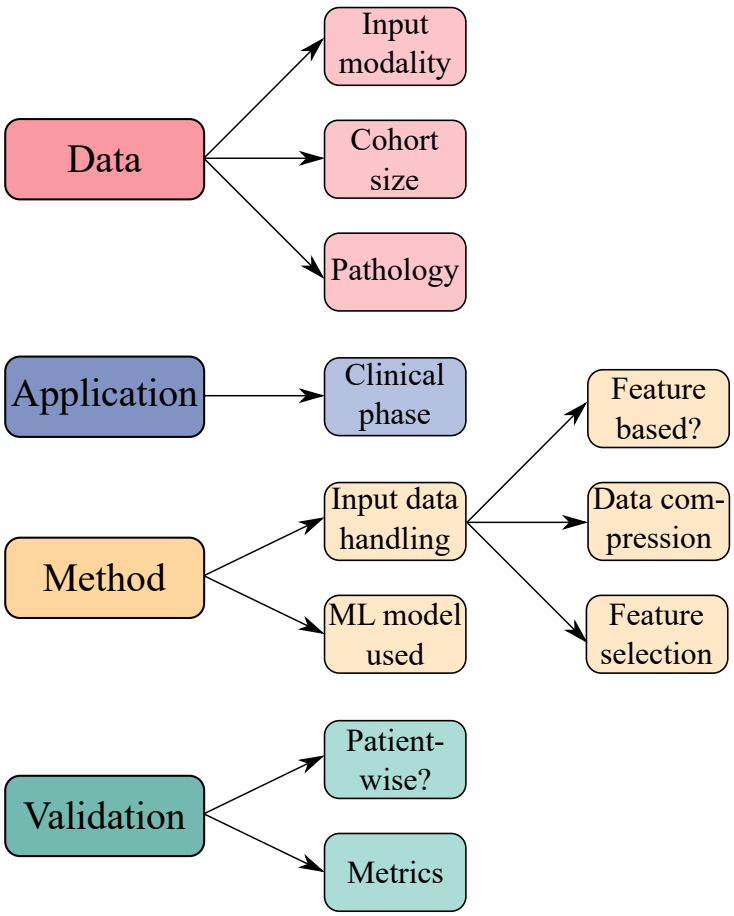


Figure 1.15 – Data was acquired on four classes: data used, clinical application, method, and validation. These classes can be composed of several items.

The class 'data' describes the cohort used to make the experiment, as well as the nature of the inputs of the ML model. We evaluated the following items:

— The input data modality type, such as imaging or MER.

— The number of patients composing the cohort.

— The pathology of the patients composing the cohort, such as PD or Essential Tremor (ET).

The class 'application' corresponds to the clinical problem addressed, by classifying at which stage of the DBS workflow this problem is encountered, using the instances 'inclusion', 'planning', 'surgery' and 'post-op', as showed in Figure 1.2.

The class 'method' describes the methodology employed to address the clinical problem, and is composed of the following items:

— The method used to handle input data, thanks to three sub-items: the eventual data compression method used, if an automatic feature selection method was employed, and if the method is feature-based, i.e. if the input data was transformed into a synthetic set of features which required a significant amount of feature engineering. For example, we did not consider as feature-based the fact of changing the domain by applying a Fourier transformation to the input data, unless specific pooling operations were done in the time domain thanks to expert domain knowledge.

— The ML model used to perform the task (ie. the classification, the regression, etc.). If several models were benchmarked, we only reported the one(s) giving the better results or the one(s) highlighted in the paper's abstract, discussion and conclusion sections.

The class 'validation' describes how the methods were evaluated, according to the following items:

— Whether or not the validation method is patient-wise. A patient-wise validation method implies that data collected from a single patient cannot be simultaneously in the training set and the testing set.

— The main metric used to evaluate the performance of the method(s). If several metrics were used, we reported the one highlighted in the paper's abstract, discussion and conclusion sections, or the one the most extensively used in the experiments.

1.3.2 Results

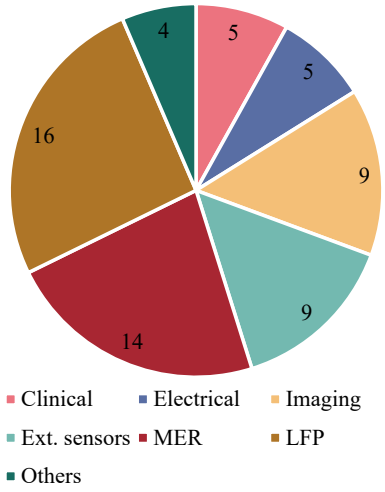
Data obtained from each study was stocked on an Excel spreadsheet. Table 1.1 presents the database collected.

Data

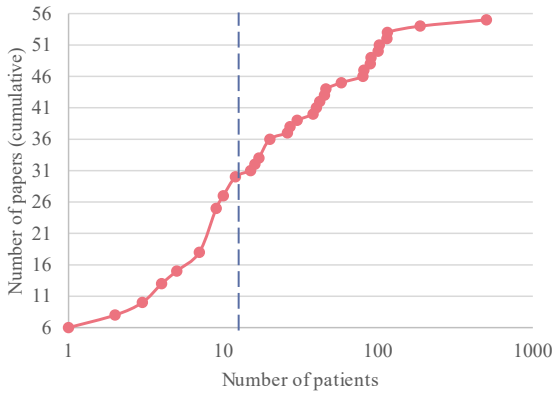
Figure 1.16a presents the distribution of the input modality type used by the model. Electrophysiological signals represent nearly half of the modality used (16 works used Local Field Potential (LFP) recorded by macro-electrode and 14 works used MER). Second, imaging data were used nine times as well as external sensors (eight times in the form of wearable sensors such as smartwatches and once with a force platform). Finally, electrical and clinical modalities were the less used with respectively five and four occurrences. Four papers used other modalities which don't belong in the previous categories (two works used Electroencephalography (EEG) and two works used electrocorticography).

First Author, Year, Ref.	Data			Application	Method		Validation	
	Input Modality	C. size	Pathology	Phase	Dim. Red.	Model	Patient wise?	Metric
Orozco, 2006 [41]	MER		PD	surgery	PCA, FB	HMM	No / N/S	acc.
Muniz, 2009 [42]	ext. sens.	45	PD	post-op.	PCA, FS	ANN	Yes	AUC
Wong, 2009 [43]	MER	27	PD	surgery	FB	unsup.	Yes	distance
Wu, 2010 [44]	LFP	1	PD	post-op.	PCA, FB	ANN	No / N/S	acc.
Guillen, 2011 [45]	MER	4	PD	surgery	FB	SVM	No / N/S	kappa
Shukla, 2012 [46]	ext. sens.	2	PD	post-op.	FB	ANN	No / N/S	acc.
Loukas, 2012 [47]	LFP	1	PD	post-op.	FB	ANN	No / N/S	acc.
Jiang, 2013 [48]	LFP	9	PD	post-op.	FB	HMM	No / N/S	acc.
Niketeghad, 2014 [49]	LFP	9	PD	post-op.	PCA	SVM, kNN	No / N/S	acc.
Connolly, 2015 [50]	LFP	15	PD	post-op.	FB, FS	SVM	No / N/S	acc.
Shamir, 2015 [51]	clin., im., elec., MER	10	PD	post-op.	FB	SVM, RF, EL, NB	No / N/S	acc.
Rajpurohit, 2015 [52]	MER	26	PD	surgery	FB, FS	kNN	Yes	acc.
Kim, 2015 [53]	im.	46	PD	planning	FB, FS	EL	Yes	mse
Khobragade, 2015 [54]	ext. sens.	1	PD	post-op.	FB	ANN	No / N/S	delay ratio
Yohanandan, 2016 [55]	ext. sens.	9	ET	post-op.	FB	RF	No / N/S	kappa
Baumgarten, 2016 [4]	elec.	10	PD	planning	none	ANN	Yes	kappa
Liu, 2016 [56]	im.	100	PD	planning	FB	RF	Yes	distance
Angeles, 2017 [57]	ext. sens.	7	PD	post-op.	FB	kNN	No / N/S	acc.
Kostoglou, 2017 [58]	clin., elec., MER	20	PD	surgery	FB, FS	RF	Yes	MCC
Houston, 2017 [59]	LFP, ECoG	1	ET	post-op.	FB	LogReg	No / N/S	acc.
Guillen, 2017 [60]	MER	5	PD	surgery	FB	ANN	No / N/S	acc.
Milletari, 2017 [61]	im.	89		planning	none	CNN	Yes	dice
Baumgarten, 2017 [5]	elec.	20	PD	planning	none	ANN	Yes	kappa
Valsky, 2017 [62]	MER	81	PD	surgery	FB, FS	SVM, HMM	Yes	distance
Mohammed, 2017 [63]	LFP	9	PD	post-op.	MRM	kNN	No / N/S	acc.
Golshan, 2018 [64]	LFP	9	PD	post-op.	PCA	SVM, EL	No / N/S	acc.
Baumgarten, 2018 [6]	elec.	30	PD	planning	none	ANN	Yes	Se/Sp
Khosravi, 2018 [65]	MER	20	PD	surgery	none	SVM	No / N/S	acc.
Lemoyne, 2018 [66]	ext. sens.	1	PD	post-op.	FB	ANN	No / N/S	acc.
Cardona, 2018 [67]	MER	5,4	PD	surgery	FB	GPR	No / N/S	acc.
Khobragade, 2018 [68]	ext. sens.	2	ET,PD	post-op.	FB	ANN	No / N/S	acc.
Oliveira, 2018 [69]	ext. sens.	38	PD	inc., post.	tSNE,FB	SVM	Yes	acc.
Shah, 2018 [70]	LFP	7	PD	post-op.	FB	LogReg	No / N/S	AUC
Yao, 2018 [71]	LFP	12	PD	post-op.	FB, FS	XGB	No / N/S	AUC
Golshan, 2018 [72]	LFP	3	PD	post-op.	PCA	SVM	No / N/S	acc.
Wang, 2018 [73]	LFP	12	PD	post-op.	FB	LDA	No / N/S	acc.
Houston, 2018 [74]	ECoG	3	ET	post-op.	none	LogReg	No / N/S	acc.
Koch, 2019 [75]	EEG	40	PD	inclusion	FB, FS	RF	Yes	acc.
Kim, 2019 [76]	im.	80	PD	planning	FB	RF	Yes	distance
Chen, 2019 [77]	LFP	12	PD	post-op.	FB, FS	SVM	No / N/S	acc.
Tan, 2019 [78]	LFP	7	ET	post-op.	FB	LogReg	No / N/S	AUC
Park, 2019 [79]	im.	102	MD	planning	none	CNN	Yes	acc.
Lemoyne, 2019 [80]	ext. sens.	1	ET	post-op.	FB	SVM, LogReg, kNN, HMM	No / N/S	acc.
Klempivr, 2019 [81]	MER	58	PD	surgery	none	CNN	No / N/S	acc.
Stuart, 2019 [82]	EEG	16	mixed	post-op.	PCA, FB, FS	SVM, RF	Yes	precision
Habets, 2019 [83]	clin.	90	PD	inclusion	none	LogReg	Yes	AUC
Camara, 2019 [84]	LFP	4	PD	post-op.	FB	SVM	No / N/S	acc.
Singer, 2019 [85]	clin., im.	114	PD	planning	FB, FS	SVM	Yes	mae
Bermudez, 2019 [86]	im.	187		planning	none	CNN	Yes	AUC
Ciecierski, 2019 [87]	MER	115	PD	surgery	FB	unsupervised	Yes	Se/Sp
Mohammed, 2020 [88]	LFP	9	PD	post-op.	FS	SVM	No / N/S	MCC
Hosny, 2020 [89]	MER	17	PD	surgery	FB	LSTM	Yes	acc.
Farrokhi, 2020 [90]	clin.	501	mixed	inclusion	FS	XGB	Yes	AUC
Valsky, 2020 [91]	MER	42	PD, dys.	surgery	FB	HMM	Yes	acc.
Baxter, 2020 [92]	im.	9	PD	planning	none	CNN	Yes	distance

Table 1.1 – Data obtained from each of the 55 papers composing the corpus. 'FB' stands for 'feature-based'. 'FS' stands for 'feature selection'.



(a) Data modality type in input of the model.



(b) Cumulative plot showing the number of patients composing the cohorts. The blue dashed line shows the median value (12 patients).

Figure 1.16 – Charts presenting the results for the 'data' class.

Figure 1.16b shows the distribution of cohort sizes. The average cohort size was 38.1 patients, with a median at 12. We counted six papers with cohorts larger than 100 patients, with a maximum at 501 and six papers that used data from a single patient.

Most of the cohorts (41 occurrences) are solely composed of PD patients. Cohorts of ET patients come second, with a total of five occurrences. Finally, five papers used a more heterogeneous cohort by mixing patients suffering from different pathologies, by mixing PD and ET patients, and/or by also studying patients suffering from dystonia or Tourette. One paper used a cohort of patients suffering from motor disorders (MD), without further specifications.

Application

Figure 1.17 presents the distribution of the clinical phase studied in the corpus. We can observe that the post-operative phase was the most extensively studied with half of the occurrences. The surgery phase and the planning phase came next with respectively 13 and 11 occurrences. Finally, the inclusion phase was the less studied one with four occurrences.

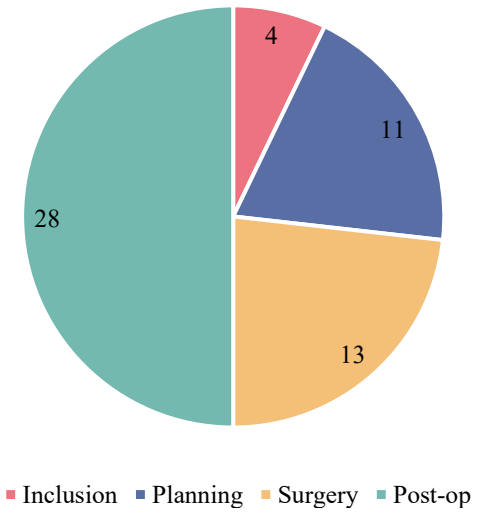
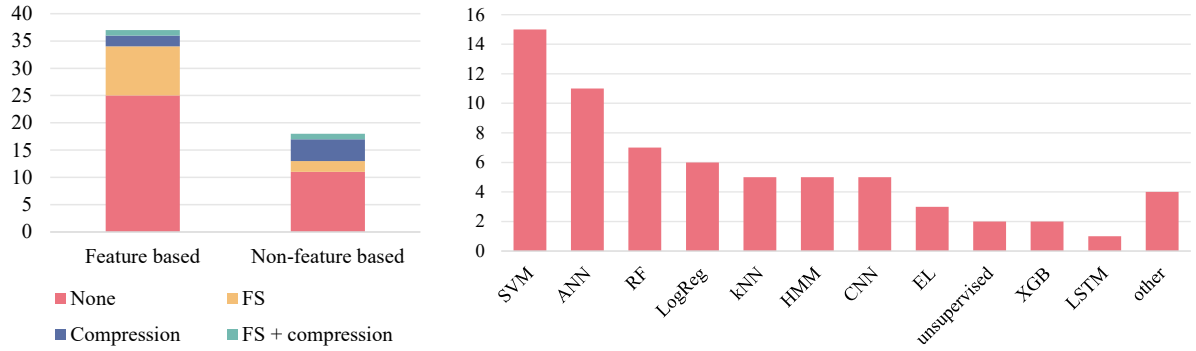


Figure 1.17 – Chart presenting the results for the 'application ' class: clinical phase studied.

Method

Figure 1.18a shows the methods used to handle the input data. The two main categories are feature-based methods and non-feature-based methods. The first category represents the contributions which required significant efforts to transform (and/or filter) the initial input space into a more readily usable form thanks to expert domain knowledge. Additionally, we noted if compression methods and/or automatic feature selection method were used. Papers were relying on feature-based methods more than 67% of the time, with 37 occurrences. In order to reduce the dimensionality of the input space, on average, feature selection or data compression techniques were used 35% of the time. The raw input space was used in 11 papers, representing only 20% of the occurrences.

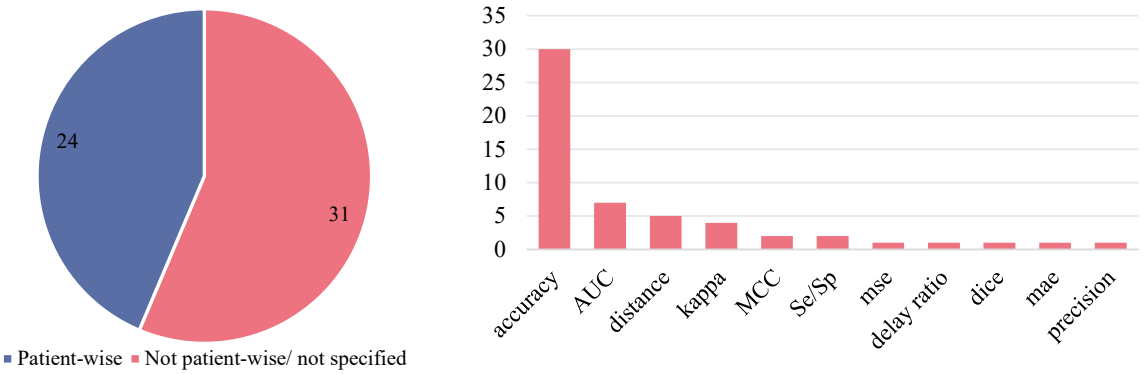
Figure 1.18b shows the model used, or the better-performing model(s) if several were benchmarked. The three mostly used models were Support Vector Machine (SVM), ANN (mostly through the form of shallow feed-forward neural networks such as Multi-Layer Perceptron (MLP)) and Random Forests (RF), with respectively 15, 11 and seven occurrences. Deeper neural networks were also used with CNNs (five occurrences) and Long Short-Term Memory (LSTM) (one occurrences). In order to study temporal sequences, Hidden Markov Models (HMM) were more regularly used than LSTM with five occurrences. Three papers used Ensemble Learning (EL) (excluding RF and Extreme Gradient Boosting (XGB)) in order to combine the predictions of several base models. k-Nearest



(a) Strategy to handle input data.

(b) Main ML model used.

Figure 1.18 – Charts presenting the results for the 'method' class.



(a) Validation strategy regarding patient splitting.

(b) Main metric used to report the performance.

Figure 1.19 – Charts presenting the results for the 'validation' class.

Neighbors (kNN) were also used with five occurrences and XGB with two occurrences. Four other models were used with one occurrence each: Naive Bayes (NB), Gaussian Process Regression (GPR), Gaussian Mixture Model (GMM) and Linear Discriminant Analysis (LDA). Finally, supervised learning represents a large majority of the papers, unsupervised models having been used only two times.

Validation

Figure 1.19a presents the validation strategy regarding the patients splitting between the training set and the testing set. A patient-wise validation strategy, employed by 24 papers, implies that data collected from a single patient cannot be simultaneously in the

training set and the testing set. 31 papers (ie. more than 56%) did not employ such a strategy, or did not specify it.

Figure 1.19b presents the main metric used to report the performances of the proposed method, and during the experiments. Classification accuracy is the most consistently used metric (30 occurrences). To report results for a classification task, Area Under the Receiver Operating Characteristic (ROC) Curve (AUC) was preferred seven times, Cohen's kappa was used four times, Matthews Correlation Coefficient (MCC) was used two times, Sensitivity and Specificity (Se/Sp) was used two times, and finally, precision was used once. To report regression performance, distance expressed in millimeters was used five times, delay ratio was used once to evaluate the prediction of timings, and mean squared error (mse) and mean absolute error (mse) were used once each. Dice was used once to report segmentation performance.

1.3.3 Discussion

Clinical problems

The most consistently studied problem with ML methods, which occurs post-operatively, is the real-time analysis of LFP signals, in the perspective of delivering adaptive DBS instead of continuous one ([44], [48], [49], [59], [63], [64], [70]–[73], [77], [78], [84], [88]). Indeed, the ability of processing and analyzing LFP signals in real time could allow designing closed-loop stimulation systems, and deliver the therapy only when needed, thus limiting the undesirable side effects and lengthening the battery duration. A complementary work was done by Loukas *et al.* [47] who proposed a complete system to record, process and display LFP signals. Houston *et al.* [74] designed a close-loop system using to cortical activity analysis, and Shukla *et al.* [46] and Khobragade *et al.* [54], [68] using surface electromyogram (sEMG) and accelerometer signals. These works, toward designing closed-loop stimulation systems, represent 19 of the 28 works focusing on the post-operative phase.

The second most common post-operative clinical problem is the analysis and quantification of the motor symptomatology of patients with external sensors (wearable sensors [55], [57], [66], [69], [80] or force platform [42]). Such automatic systems can be precious for clinicians by offering them an objective, automatic and quick feedback for therapies in order, for example to compare several therapy tuning and combinations. Indeed, the therapy tuning scope, between stimulation parameters and drug treatment

dosage, is large and can't be assessed exhaustively, which can lead the clinician to choose a sub-optimal configuration. Other works have been done toward automatizing this post-operative phase: Connolly *et al.* [50] proposed a system predicting which contact is optimal from LFP recordings. Shamir *et al.* [51] proposed a predictive system using inputs from several modalities (clinical, therapy (medication and stimulation) and demographic data) in order to narrow the research space both for medication dosage and stimulation parameters. Lastly, Stuart *et al.* [82] used EEG to predict effective stimulation in real time.

The second most consistently investigated phase is the surgery itself, the target location being the clinical problem. Every paper addressing this problem used MER analysis ([41], [45], [52], [60], [62], [65], [67], [87], [91]). Complementarily, Wong *et al.* [43] proposed a method to project MER in a 2D plan in order for clinicians an alternative way to visualize and interpret it. MER being frequently polluted by artifacts, Klempivr *et al.* [81] and Hosny *et al.* [89] proposed systems to detect them, in a perspective of curating the signals for downstream analysis. Lastly, Kostoglou *et al.* [58] used MER features, a few clinical scores, demographics and contact location to predict the clinical improvement that could result from the stimulation of various locations, in a perspective of placing the electrode based on functional criteria rather than anatomical ones. As MERs are electrophysiological signals, these papers are similar in methodology to the ones using LFPs.

The planning phase comes third, the clinical problem being how to choose the stimulation targets and electrode trajectories before the operation. A first strategy is to assist the surgeon by automatically segmenting the subcortical structures of interest from pre-operative images ([53], [56], [61], [76], [79], [92]). Such works are precious because the surgical targets can be small and not well visible imaging scans. An alternative strategy is to propose to the surgeon functional criteria instead of anatomical ones for the choice of a stimulation site. Baumgarten *et al.* [4]–[6] and Bermudez *et al.* [86] proposed clinical efficacy probability maps to visualize the expected clinical effects of stimulation of several locations around the structure of interest. Singer *et al.* [85] went even further in this idea by predicting the optimal electrode location directly.

Finally, the inclusion phase is the most rarely studied one. Oliveira *et al.* [69] proposed a method to visualize on a two-dimension space the motor symptomatology of the patient from electromyography sensors in order to ease clinicians' interpretation of the patients' motor scores. Habets *et al.* [83] proposed a predictive system to identify weak

motor responders from pre-operative clinical data and demographics for patient screening purposes. In the same screening assisting tool objectives, Koch *et al.* [75] proposed a system to classify patients regarding their cognition from EEG. Lastly, Farrokhi *et al.* [90] attempted to find factors of surgery adverse outcomes, such as infections or hemorrhages.

A great variety of model used

When it comes to the choice of a ML model, we can see that the most widely used one is SVM. It is not surprising as SVMs are widely known as well performing, simple to train for both classification and regression problems, even for small databases. Among the papers comparing several models, SVMs were on the top-performing ones [80], [82]. Shallow feed-forward ANNs come second, probably because of the recent trendiness around deep-learned ANNs. Deeper and more specialized ANN structures were also used. First, CNNs were used four times for image analysis: three times for subcortical structures segmentation with the VGG model [79], a modified ResNet structure [86] and a custom structure called Hough-CNN [61] based on Hough voting, and once for MER spectrogram analysis for artifact detection with the AlexNet model [81]. Second, RNNs were used once with LSTM for MER artifact detection [89]. We can also mention the usage of an interesting paradigm called EL, which consists in using several base models to make a prediction. Bagging was used eight times with RFs, and once by Kim *et al.* [53] making it the third mostly used model. Boosting was used twice with the XGB model, and stacking was used by Golshan *et al.* [64] and Shamir *et al.* [51].

How to handle complex and high-dimensional inputs? The prominent role of feature engineering

An important limit of ML studies is the curse of dimensionality: the greater the input dimensionality is, the exponentially greater the number of training samples are required to guarantee that data points are not too sparse in the input space. In DBS, the number of training samples are usually limited. Therefore, limiting the dimensionality of the input space to ease the training process can be an interesting strategy, even if it comes at the cost of the shrinkage of the amount of information passed to the ML model.

Two common strategies can be to unsupervisedly compress the data and/or to automatically select the features, but such approaches are not majoritarian. The most common strategy is to transform the original, raw input space into a set of features thanks to expert knowledge of the domain. For example, Kostoglou *et al.* [58] synthesized the MER

signals by computing domain-specific features such as the mean inter-spike interval, or the power band ratio of the signal in different pre-determined frequency bands.

A minority of papers chose a fully data-driven method by feeding the model with raw data straightforwardly, without reducing it in the form of features, compressing it or automatically selecting some of them. Naturally, the four works using CNNs fall under this umbrella. Indeed, the point of a CNN is to take advantage of the spatial inter-correlations on the input images by applying a series of learnable convolution kernels on it. Therefore, CNNs learn an optimal dimensionality reduction strategy in a supervised manner. Among the other occurrences are the three papers of Baumgarten *et al.* [4]–[6] and the paper of Habets*et al.* [83], because the number of inputs is low (respectively four stimulation parameters and 15 clinical scores and demographics).

We can interestingly mention two papers which compared a feature-based method with a fully data-driven method. First, Khosravi *et al.* [65] compared the utilization of state-of-the-art features for MER analysis for the location of the STN versus the utilization of the Fourier coefficients of the raw signal. Second, Baumgarten *et al.* [5] compared two methods for predicting the occurrence of PTSE during the stimulation: a VTA-based method versus a method using the raw information straightforwardly, which is the three-dimensional location of the contact and the stimulation voltage. Both papers showed a superiority of the fully data-driven paradigm which outperformed the feature-engineering approaches.

Inter-patient variability: the elephant in the room

Small cohorts are problematic in ML because they limit the performance of predictive systems (several papers [64], [77], [82], [88], [90] stated lack of data as a limitation). Furthermore, the pathologies treated by DBS are heterogeneous causing a high inter-patient variability, on top of intra-patient variability (because the clinical state of the patients can fluctuate or because the recording conditions may vary). Therefore, a system trained on one patient or on one recording configuration is not likely to have good performance on another one, or later in time, which interrogates regarding its prospective usability.

The contribution of Khobragade *et al.* [68] is a great illustration of this phenomenon. They gathered surface electromyography and accelerometer data from two patients through several trials spread on different sessions, with at least a week between consecutive sessions. They did two experiments: the first one by training one model per patient and per session, therefore testing the model on trials of the same sessions. For the second

one, they trained one model per patient, but trained the model on a set of sessions and tested it on other sessions. They obtained a perfect accuracy for the first experiment, but the median performance dropped to 46.15% for the second one. On the same extent, Rajpurohit *et al.* [52] reported results with a patient-specific feature normalization scheme (therefore not applicable prospectively), and with a patient-independent normalization scheme. Not surprisingly, the classification error rates of the patient-specific scheme are much lower (in the range of 0.0711 to 0.1353) than the patient-independent scheme error rates (in the range of 0.102 to 0.1979).

A majority of papers reported performance on a single patient, chose to train one model per patient and validated it on the same patient, trained a single model evaluated with a non-patient-wise validation method, or did not give this information ([41], [44]–[51], [54], [55], [57], [59], [60], [63]–[68], [70]–[74], [77], [78], [80], [81], [84], [88]). The results reported by these contributions are interesting as preliminary work, but cannot safely be considered as representative of a real, prospective usage. Oppositely, some reported performance by employing a patient-wise validation method ([4]–[6], [42], [43], [52], [53], [56], [58], [61], [62], [69], [75], [76], [79], [82], [83], [85]–[87], [89]–[92]). While it does not ensure the complete absence of data leakage (that could occur, for example, by mixing the validation and testing sets, or by selecting or normalize features with the whole database), these results can be considered as more reliable, and more representative of prospective-usage performance. Inter-patient variability remains an open problem for most of the contributions and is likely only solvable thanks to extensive data collection. Several papers stated that the lack of variability in the cohort limits the generalizability of the results ([51], [75]), that further validation has to be done on other recording conditions [62] or that inter-patient variability was a limiting problem [74].

A reproducibility and comparison problem

Another recurrent problem is the disparity in the validation methods which keeps from comparing the results of different contributions together.

As an example, all of the following contributions report the results of the location of the STN with MER but differ in the validation process:

— Guillén-Rondon *et al.* [45], [60] mixed patients and different portions of the same signals in training and test sets, which is an identified source of data leakage, leading to likely overestimated performance.

— Khosrava *et al.* [65] didn't specify if the validation was done on a separated set of patients.

— Rajpurohit *et al.* [52] used a leave-one-patient-one validation strategy.

While the contributions on the same topic are proliferating, it is almost impossible to draw a hierarchy out of the methods proposed because the results reported do not carry the same meaning.

On the same extent, methods are not easily reproducible for benchmarking purposes as no paper but one mentioned having publicly shared their code or their data (only Oliveira *et al.* [69] publicly shared the features computed on their database).

1.3.4 Conclusion

We conducted, to the best of our knowledge, the first systematic review on ML for DBS, and identified several limits from the analysis of a corpus of 55 papers:

First, we have seen that only a few studies concern the inclusion phase. To us, there is a real opportunity for ML here because, as there exist several clinical challenges and a lot of complex and high-dimensional data arising from several pre-operative modalities (such as clinical testing, patient questionnaires, imaging data or demographics). These factors make this phase challenging to address but also prompt for data-driven methods providing that adequate methods are employed and large enough databases are available.

Methodologically speaking, the majority of studies uses simple models with either a few features as input, aggressive dimensionality reduction methods (with automatic feature selection or feature-engineering). While it ensures the input data to be readily usable for simple models, it also limits drastically the information that can be leveraged by the model, and therefore limits the performance and the practicable complexity of the prediction task. We think more ambitious problems, or higher levels of performance could be achieved by employing more bottom-up, data-driven paradigm. Two papers compared a feature-oriented method to a data-driven one and showed better results for the latter approach, supporting this hypothesis.

Third, small cohorts are often used: a lot of studies collect data that are not usually collected in the clinical routine, thus limiting the number of patients. Small cohorts, first, are problematic because they limit the generalization performance of the model and impose limits regarding the dimensionality of the input, because of the curse of dimensionality. Some paper overcame this problem by employing data augmentation when

possible (for example, by splitting a ten seconds LFP signal into several two seconds ones). Unfortunately, this strategy does not address the second problem caused by small cohorts, which is that they cannot cover the great heterogeneity of the studied populations.

Additionally, a lot of contributions validated their method in a non-patient-wise manner.

Both factors make the performances reported by a lot of works likely not representative of a prospective use of the system.

Lastly, to us, better efforts have to be done to give maximum details concerning the validation method employed, and if possible to share the code and the data to ease the reproducibility. We showed indeed that a lot of papers are sharing similar, if not identical objectives but no paper benchmarked different approaches, or compared the performance of their method to methods proposed in the literature, as it would imply to reproduce the methods to evaluate them with the same database, and by employing the same validation method. We think this area of research would benefit from more uniform validation methods and explicit validation guidelines, and further work would be required in this direction.

Section conclusion

ML has been heavily used in DBS since 2014. From a systematic review, we identified several points to be considered for future work. First, most of the contributions are centered around feature engineering. The input space is often limited to a few components, or features are designed to drastically synthesize the raw data into a more readily usable form thanks to domain knowledge. Although this approach ensures good performance of the downstream ML model by addressing the curse of dimensionality, two papers showed that better performance could be achieved by more data-driven approaches. Second, the inter-subject variability remains an open problem and is often neglected or passed over in silence. We think, for retrospective studies, greater efforts should be made on data collection and evaluation methods should be done on unseen patients, in order to be representative of a prospective usage. Last, efforts have to be done to ease method reproducibility and make method benchmark possible. When possible, contributions should compare their method to a known baseline to give a clearer idea on the interest of the methods proposed.

1.4 Contributions

On Section 1.1, we have seen that DBS is a very effective but complex procedure composed of different phases. We have isolated several problems and challenges faced by clinicians. These problems are difficult to solve from an engineering point of view because they involve many data modalities presenting their own difficulties to be processed. For this reason, ML looks like an appropriate methodology [38] and have been applied on a variety of DBS problems over the past decade. Indeed, ML models are powerful tools able to find inter-correlations on the data by learning from a retrospective database, and can be used prospectively providing that the training phase was successful. A systematic review allowed to understand the trends and the limits of ML for DBS. The most important conclusion point is that a majority of the contributions were heavily centered toward feature engineering. On the other hand, pure data-driven approaches, i.e. pipelines not relying on expert knowledge seem to be comparatively more interesting.

In this thesis, we developed two non-knowledge-based CDSSs to address two clinical problems identified in Figure 1.2 where we spotted an opportunity for data-driven ML to be appropriate and powerful. We proposed four contributions, each of them having in common to be bottom-up oriented, without having to rely on feature engineering nor on strong clinical or engineering priors.

Our first CDSS addresses the prediction of DBS clinical outcomes from pre-operative biomarkers in inclusion phase. The heterogeneity of PD is known, and the clinical outcomes of DBS is a function of the state of advancement, the form and the particularities of the disease of the patient [93]. For this reason, an important step to predict the clinical outcomes of DBS is to represent this heterogeneity efficiently, in order to feed the predictive system with meaningful and accurate bio-markers. A common approach is to define subtypes of PD explicitly, but these subtypes are poorly reproducible [93]. This observation motivated us to propose two contributions to synthesize raw clinical information, respectively answers to clinical questionnaires and T1-MRI, to readily usable and informative forms, in a full data-driven fashion. Finally, another contribution presents and validates the usage of both of these methods with a downstream predictive system. Here are more details concerning these three contributions:

1. Firstly, we proposed a deeply-learned autoencoder to compress and curate clinical tests and patient questionnaires. Answers to these questionnaires are highly informative regarding the particularities of the disease of the patient, making them precious

in input of our predictive system, but are suffering from being high-dimensional and from their propensity of suffering from missing data. We prove that our method is able to solve both problems better than linear baselines, and investigated on its ability to impute missing data on various scenarii, making clinical data more readily usable for downstream applications. This work is presented in Chapter 2, and is under minor review in Artificial Intelligence in Medicine (Elsevier).

2. Secondly, we proposed a ML-based pipeline to unsupervisedly compress the displacement field of the bilateral striatum (composed of the bilateral putamen and caudate) from T1-weighted MRI. We prove that striatal shape displacements are very informative when exploited by appropriate ML methods, making them valuable as staging biomarkers. Notably, we showed that striatal displacements are more relevant than motor symptomatology to diagnose prodromal stages of the disease. This work allows a novel and powerful way of drawing information from the most consistently used scan in DBS for PD. It is presented in Chapter 3 and has been published in NeuroImage: Clinical (Elsevier).
3. Thirdly, we proposed a ML-based workflow that takes as input clinical and morphological information with the methods presented in the previous two points, as well as demographics. We proved this workflow to be able to predict most of the post-operative clinical scores at 3, 6, 12 and 36 months after surgery. This work represents an innovative way of addressing the DBS clinical outcome prediction problem, and is innovative as it considers several modalities simultaneously. Moreover, we designed this workflow to be flexible to warranty that it can be tested on various clinical sites and with various biomarkers. This work is presented in Chapter 4 and has been submitted to Transaction on BioMedical Engineering (IEEE).

The second CDSS addresses the location of the STN using MER, thanks to a new CNN structure, with one contribution. This work alleviates the limits raised in the previous section by exploiting raw data directly without simplifying the signal with feature engineering. We showed that one second of signal is enough to make reliable predictions, and that our proposed CNN structure is superior to standard computer vision CNN ones. This work represents a promising new way of automatizing and speeding up STN border detection, and is presented in Chapter 5. This work has been accepted at the 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society in conjunction with the 43rd Annual Conference of the Canadian Medical and Biological Engineering Society.

Chapter conclusion

In the context of neurosciences, where the links between input data and desired tasks are difficult to find and where big databases are complicated to gather, the question of the relevance of bottom-up, highly data-driven approaches remains open. To explore this new, ambitious data-science paradigm, in the spotlights since 2012 with the rise of DL, we propose four contributions to address two clinical problems in DBS from a new angle. These four contributions will be presented in the next chapters.

Chapter 2

PatiNAE: Patient clinical data normalization using auto-encoders

This chapter is the first in a series of three towards the construction of a patient-specific CDSS for the prediction of DBS clinical outcomes from pre-operative information, in order to assist clinicians during clinical *staff*.

In order to be considered as patient-specific, a model has to make its predictions based on inputs representative of the clinical state of a patient. Some of the most crucial sources of information are the answers of medical questionnaires and exams collected from the patients during their visits at the hospital. These clinical tests are performed to inform the clinicians about the state of advancement, the form and the particularities of the patient's disease, making them an important source of information for any patient-specific prediction tool.

It is however difficult to leverage this information as:

— the patient's neurological functions are encoded in several variables (e.g. the motor symptomatology is tested by several items of several tests),

— many variables can be affected by different neurological functions,

— variables are noisy (e.g. because a patient can be in a 'bad day' when performing certain tests),

— the test's answer databases generally present a lot of missing values and missing tests (e.g. because some tests are not realized for every patient at every visit), and

— the tests are numerous, increasing the dimensionality of the answer vectors.

In this chapter, we propose a method which bundles both data compression and imputation to address these last two problems for the purpose of making clinical testing data more readily usable for our patient-specific prediction tool proposed in Chapter 4.

This chapter is on minor revision at Artificial Intelligence in Medicine [1].

1. Peralta, M., Haegelen, C., Jannin, P. and Baxter, J.S.H. (2020). Data Imputation and Compression For Parkinson's Disease Clinical Questionnaires. Artificial Intelligence in Medicine (minor revision).

2.1 Abstract

Medical questionnaires are a valuable source of information but are often difficult to analyse due to both their size and the high possibility of having missing values. This is a problematic issue in biomedical data science as it may complicate how individual questionnaire data is represented for statistical or machine learning analysis. In this paper, we propose a deeply-learnt residual autoencoder to simultaneously perform non-linear data imputation and dimensionality reduction. We present an extensive analysis of the dynamics of the performances of this autoencoder regarding the compression rate and the proportion of missing values. This method is evaluated on motor and non-motor clinical questionnaires of the Parkinson's Progression Markers Initiative (PPMI) database and consistently outperforms linear coupled imputation and reduction approaches.

2.2 Introduction

Data representation is a critical problem in biomedical data science, in which the data available concerning an individual patient can be simultaneously large, uncertain, heterogeneous, and incomplete. It is common knowledge in the machine learning community that accurately curating data and providing a uniform representation are often critical to the success of later approaches. Medical questionnaires are one of the major ways of assessing the clinical state of a patient, providing crucial information necessary to diagnose and monitor patients. Extracting this information is crucial to ensuring patient care in the era of computerised, personalised medicine. Unfortunately, medical questionnaires databases often suffer from missing values for various reasons: some tests cannot be performed by some patients or aren't performed at each visit, problems can occur when data is computerised, or paper records can be lost. This issue is well known to the research community and addressing it remains an active field of research [94] [95].

Data imputation is an element that addresses this type of heterogeneity by estimating the value of missing values in an incomplete data vector based on the non-missing values and the population distribution [96]. Data imputation methods are often employed in research to estimate the value of missing data allowing for downstream analysis or statistical processing to be performed. It has been shown that ignoring data points with missing values not only substantially limits the performance of downstream analysis [97], but also that more accurately imputed data yield improves performance in downstream analysis [98] [99].

A related problem is dimensionality reduction in which a large data vector is reduced into a smaller one that can be more readily analysed to monitor and assess the patient's situation. Keeping a high dimensional input data vector leads to a problem known in the artificial intelligence community as the 'curse of dimensionality' [95]. It has been shown, specifically on electronic health records data, that projecting data into a smaller latent space yield better results than using the original space [100]. The current clinical paradigm is to aggregate related values, summing them into a single score and normalising to account for any missing values. Although simple to implement, this method has been difficult to design [101] with the exact partitioning and summation schemes leading to model misfit and misrepresentation in questionnaires used in Parkinson's disease [102].

Separately, both data imputation and dimensionality reduction are well-known issues in health informatics recently tackled by machine learning [95], and have both been proven

to be beneficial for downstream analysis. Abedia *et al.* [103] showed that projecting incomplete data to a compressed latent space can be beneficial for data imputation, motivating us to tackle and evaluate these problems altogether. Although compressive AE and DL have already been used for data imputation [104], there is, to the best of our knowledge, no paper investigating the dynamics of imputation and reconstruction performance varying the compression amount and the proportion of missing values.

AE are a family of ANN trained to reproduce its input with a lower dimensional immediate stage or bottleneck [105], which have long been used for data imputation [106]. These networks often consist of a series of encoding layers leading up to a central bottleneck which is then followed by a symmetric series of decoding layers. The central bottleneck creates an *internal representation* of the input data with a lower dimensionality. AE-based approaches to analysing medical data have been shown to provide useful patient representations for screening broad disease classes [107].

In this chapter, we bundle both prospective data imputation and dimensionality reduction into a single method, allowing us to effectively summarise the PD patient's questionnaire data in a way that can be more readily used for further machine learning methods and population research. We also investigate the impact of differing levels of compression and differing levels of data corruption on imputation performance.

2.3 Theory and Related Work

2.3.1 Data Imputation

Imputation can be conceptually split into methods that are applied *prospectively*, where a possibly complete or incomplete training database is used to estimate missing values for an incomplete and previously unseen data vector, and methods that are applied *retrospectively*, where information from an incomplete database is extracted in order to estimate its own missing values. In a clinical context, prospective imputation is of greater utility, allowing new patient records to be processed, although retrospective is more commonly used in research contexts in which an entire database is often analysed at the same time.

The most common way of performing retrospective imputation is case deletion, in which every sample with at least one missing value is removed from the database. This method has the benefit of being easy to use, but suffers from two main drawbacks, in addition to not being capable of prospective use. First, if missing values are distributed amongst a large number of samples, it can substantially reduce the size of the database, limiting the power of any statistical or machine learning method. Secondly, if the probability of which values are missing is not independent or changes based on the value the variable would have otherwise taken, removing incomplete lines can introduce bias into the study [108].

A common but more nuanced method for data imputation is fitting a linear model to the data. Principal Component Analysis (PCA) in particular can be naturally extended to perform prospective imputation by removing the PCA eigenvector components corresponding to the missing values when calculating the PCA scores but using the full eigenvector when transforming the scores back into the data space. Assuming the training database also contains missing values, the PCA decomposition can be determined through several methods [109]. Pairwise correlation PCA (PPCA), for example, computes the mean vector and correlation matrix from all the data vectors available with the corresponding values [109]. Iterative PCA (IPCA) is an expectation-maximisation algorithm that iterates a process of PCA decomposition and imputation until the PCA decomposition converges [110], [111]. These methods are designed to preserve the mean and covariance of the observed data through the process of imputation. The fact that PCA is also a common dimensionality reduction method makes it even more suitable for patient normalisation, although their linear nature may be problematic as aspects of the question may be coupled in a fundamentally non-linear manner.

2.3.2 Autoencoders

To perform non-linear dimensionality reduction, AEs have shown promising results. They have the benefit of capturing more complex or non-linear relations between the inputs. Stacked denoising AEs [112] are particularly useful as they have highly robust denoising capabilities resulting from having noise (theoretically of the same distribution as would be observed in the testing phase) injected into their input during training.

AEs can also be designed with large-scale data imputation explicitly in mind. For example, *correlation neural networks* [113] attempt to find correlated, modality independent internal representations from missing modality problems, where large portions of the data vector are missing simultaneously. Unfortunately, the regularisation term which encourages this correlation requires a low number of modalities and complete data vectors for training which limits their applicability to medical questionnaires where this may not be the case.

Nevertheless, ANN can be difficult to design and to train for various reasons. Primarily, the shape of the neural network highly affects its performance. For example, even assuming low parameterisation and risk of overfitting, a shallow neural network may not be able to capture the non-linearities of the inputs, while a too deep neural network could suffer from training issues such as inability to propagate gradients effectively.

In this paper, we propose a method relying on an autoencoder architecture with a densely-connected encoder and decoder consisting of fully-connected layers, which we call a Fully Connected Autoencoder (FCAE), to address both data imputation and dimensionality reduction.

2.4 Materials and Methods

The input data is first rescaled in the range [0,1], with min-max scaling based on the highest and lowest values possible for each question. Although this does not necessarily reflect the semantic meaning behind different levels in the ordinal variables, it does ensure that the number of levels in each does not excessively bias the training of the network and is commonly used even for ordinal data [114]. Denoising AE are then constructed with masking noise applied to the input layer in order to simulate missing values. Unlike truly missing values, their reconstruction quality can still appear in the loss function. This is implemented through a custom masking layer which randomly removes an entire modality (collection of inputs taken from the same test and thus tend to be missing or be present *en bloc*).

The input to the encoder consists of both the data and a mask which indicates the presence of missing data. The encoder consists of the bias layer, which chooses particular initial values to assign to missing variables as a form of initial naive imputation, followed by a series of dense layers, the output of which is concatenated to the input for the successive layers, as shown in Figure 2.1(a). This alternation between dense and concatenation layers is motivated in a similar way as residual networks [115] in that they allow for short-cuts, minimising issues with propagating gradients while allowing for higher depths to be used to capture non-linearities, and has already been used successfully for data imputation [116]. The final layer of the encoder is a dense layer used to estimate the internal representation. The decoder is constructed similarly to the encoder, with a series of computational blocks each receiving the internal representation and the output of all previous decoder layers as input. Each dense layer is composed of 10 neurons and both the encoder and decoder have a depth of 7 layers, leading to an overall depth of 15. Dropout (5%) was applied to the input of each layer except the final one. Rectified Linear Units (ReLU) were used as the activation functions for each dense layer. The proposed structure (shown in Figure 2.1) is similar to that of *multimodal AE* [117] with the exception of our concatenation structure and that the noise operator is performed within the network as a layer rather than used to augment the dataset prior to training.

In order to improve the convergence of the autoencoder during training and to minimize the effects of random weight initialisation, each FCAE began with an initialisation step. The bias layer was initialised to replace missing values with the mean value of the respective variable. Each encoder layer was then greedily initialised to the PCA transform

that preserved the largest amount of information from its input. Each decoder layer was initialised with linear regression to create the optimal reconstruction given the input to that particular output layer. This initialization guarantees that the training set performance of the FCAE is at worst equivalent to that of the pairwise PCA variant. This step was performed using the entirety of the training dataset simultaneously, but did not involve data augmentation.

The hyperparameters for the network were determined in a two-step process. First, we defined the topology of the network using a grid search, in order to optimize the number of neurons per dense layer and the depth of the network while keeping the other parameters at a constant value. Second, we optimized training-related parameters (keeping the network structure fixed), such as learning rate, batch size and dropout rate, in a Bayesian manner using a Gaussian process as a surrogate model and expected improvement as the criterion.

The networks were implemented in Keras using a TensorFlow back-end with NAdam as the optimizer. We have made available the code to build and compile our proposed autoencoder, at `https://github.com/m-prl/PatiNAE`.

2.4.1 Accuracy, Loss and Regularisation Metrics

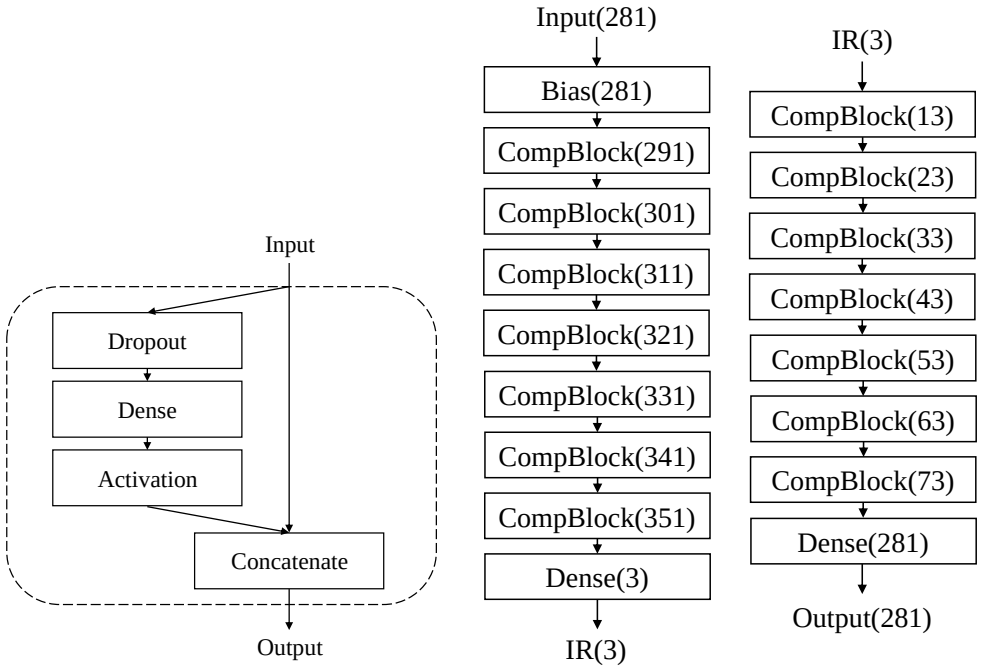
For the purpose of evaluating each compressing data imputation approach, two measures of accuracy based on the input, x, and reconstruction values, $\hat{x}$, were used:

$$A_1 = \frac{1}{K} \sum_{i=1}^{N} \sum_{i=j}^{M} (\hat{x}_j^{(i)} - x_j^{(i)})^2 * M_j^{(i)} \tag{2.1}$$

$$A_2 = \frac{1}{U} \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{x}_j^{(i)} - x_j^{(i)})^2 * (1 - M_j^{(i)}) \tag{2.2}$$

where K is the number of known values and U the number of unknown values in the dataset being evaluated, and $M_j^{(i)}$ is a mask identifying the known values. For A_2, additional unknown values must be inserted into the dataset in order for their ground-truth value ($x_j^{(i)}$) to be known for evaluation purposes.

For training, the loss metric, analog to the one used by Sanchez *et al.* [98], used a weighted mean squared reconstruction error using both a binary mask to identify missing data (M_{miss}) and a second binary mask to identify data that has been dropped in the



(a) Computational Block (b) Encoder (c) Decoder

Figure 2.1 – Structure of the FCAE, relying on the chaining of a residual substructure called 'Computational Block', presented in (a). The input is passed through the encoder, which produces the internal representation, as shown in (b). The decoder tries to reconstruct/impute the input from this internal representation, as shown in (b). The size of the output of each block is shown in parenthesis. In the given example, the internal representation (IR) size is equal to 3.

process of data augmentation (M_{drop}):

$$L = \frac{1}{K+U}\sum_{i=1}^{N}\left(\sum_{i=j}^{M}(\hat{x}_j^{(i)} - x_j^{(i)})^2 * (1 - M_{miss,j}^{(i)}) \quad + 4 * \sum_{i=j}^{M}(\hat{x}_j^{(i)} - x_j^{(i)})^2 * M_{drop,j}^{(i)}\right) \quad (2.3)$$

The weighting factor of is equivalent to put a higher importance into reconstructing the dropped data and thus preferentially improves the network's A_2 accuracy. A weighting factor of 4 showed the most interesting results.

2.4.2 PPMI Questionnaire Database

The Parkinson's Progression Markers Initiative (PPMI) [118] is a program sponsored by the Michael J. Fox Foundation for Parkinson's Research. It is an observational clinical study which tracks cohorts of subjects with different forms of Parkinson's disease for up to

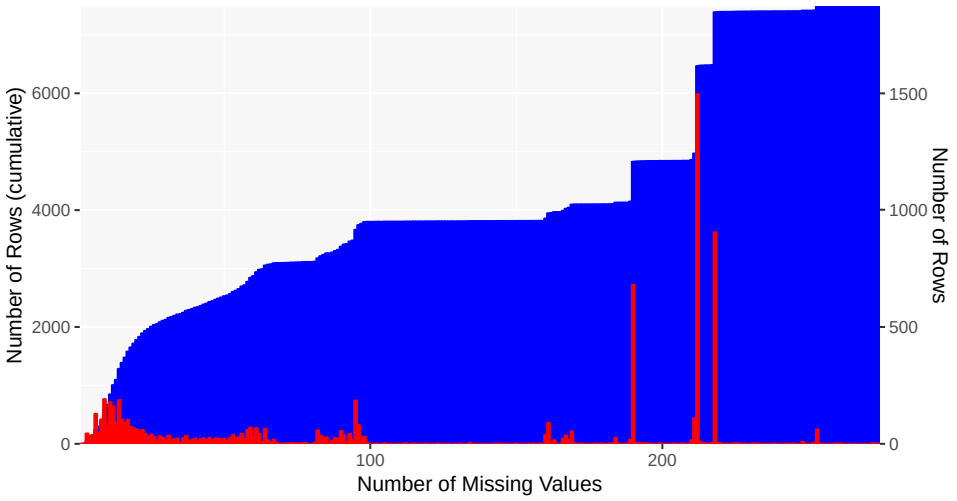


Figure 2.2 – Histogram of PPMI questionnaire data by number of missing values. The cumulative histogram is shown in blue (left axis) and the frequency in red (right axis).

3.1 SPEECH **SCORE**

Instructions to examiner: Listen to the patient's free-flowing speech and engage in conversation if necessary. Suggested topics: ask about the patient's work, hobbies, exercise, or how he got to the doctor's office. Evaluate volume, modulation (prosody) and clarity, including slurring, palilalia (repetition of syllables) and tachyphemia (rapid speech, running syllables together).

0: Normal: No speech problems.

1: Slight: Loss of modulation, diction or volume, but still all words easy to understand.

2: Mild: Loss of modulation, diction, or volume, with a few words unclear, but the overall sentences easy to follow.

3: Moderate: Speech is difficult to understand to the point that some, but not most, sentences are poorly understood.

4: Severe: Most speech is difficult to understand or unintelligible.

Figure 2.3 – First question of Part 3 (motor examination) of the MDS-UPDRS test (version as of May 2, 2019).

Questionnaire	% comp.	% incomp.	% mod. miss.	miss. rate
UPDRS I	98.84	0.03	1.13	16.67
UPDRS I PQ	98.80	0.04	1.16	23.81
UPDRS II	96.52	2.43	1.05	10.99
UPDRS III	2	94.73	3.27	6.51
UPDRS IV	55.91	0.04	44.05	61.11
S&E	98.79	0	1.21	N/A
PASE	13.72	15.41	70.87	22.12
SCOPA AUT	0	50.89	49.11	9.37
MCI	38.65	0	61.35	N/A
Ger. Dep.	50.68	0.17	49.15	9.23
QUIP	32.03	18.85	49.12	15.04
STA Inv.	50.43	0.41	49.16	9.27
Benton JLO	40.60	0	59.40	N/A
Hopkins VLT	40.76	0.81	58.42	39.47
L-N Seq. PD	4.15	37.25	58.60	30.60
MoCA	46.33	0.12	53.55	13.10
Sem. Flu.	41.54	0	58.46	N/A
Symb. Dig. Mod.	41.44	0.07	58.49	50.00
Epworth sleep.	50.80	0.04	49.16	12.50
REM SD Quest.	54.07	0.65	45.27	6.32
Average	47.80	11.10	41.10	21.01

Table 2.1 – Statistics of PPMI questionnaires regarding missing values. The second, third and fourth columns show the percentage of rows being complete, with missing values, and with the entire modality missing, respectively. The last row shows the mean percentage of missing values for the incomplete rows.

8 years with the goal of identifying biomarkers of disease progression using MR imaging, biologic sampling as well as clinical and behavioural assessments.

This work is focused on the clinical and behavioural assessments which were designed to provide a satisfactory amount of information regarding the patient's motor, cognitive, neuro-behavioural and neuro-psychological state. The database was built collecting and merging data from most of the motor and non-motor tests. The tests presented on the database are: MDS-UPDRS (all parts), Physical Activity Scale for the Elderly - Household Activity, Modified Schwab & England ADL, SCOPA-AUT, Clinical Cognitive Categorisation, Geriatric Depression Scale (Short), Questionnaire for Impulsive-Compulsive Disorders (QUIP), State-Trait Anxiety Inventory, Benton Judgement of Line Orientation, Hopkins Verbal Learning Test, Letter-Number Sequencing (PD), Montreal Cognitive Assessment (MoCA), Semantic Fluency, Symbol Digit Modalities Text, Epworth Sleepiness Scale, Features of REM Behaviour Disorder. Some parts were discarded because they were performed too sparsely (on too few patients or just at one visit). We built our database by pooling the tests of each cohorts at each visit, on April 2018. The constructed PPMI database has 281 columns and 7490 rows across 1011 patients. 42.9% of the data is missing in a highly heterogeneous manner. The questions used in these tests mostly permit ordinal answers. These tests do not include questions that permit categorical answers, but there are a few that permit continuous-valued ones. Figure 2.3 shows the first item of part 3 of the MDS-UPDRS test, which is a question with an ordinal answer. Table 2.1 gives statistics regarding missing values and modalities for each questionnaire used in this study. The names of the questionnaires have been shortened for clarity purpose, and are in the same order than presented in this section. This table shows a great heterogeneity in the way that values are missing. Nonetheless, the major cause of missing values is when the whole modality is missing *en bloc*. Note that the large majority of missing values are due to the protocol design, as not all tests are performed, nor are intended to be performed, at each visit for each cohort. This is well described in the PPMI database's protocol information.

An extensive study on the MDS-UPDRS questionnaire (present in the PPMI database) as been performed by Goetz et. al. [119], showing that the loss of information is consequential even with only a few missing values. In their analysis, removing 0-27% of answers completely at random drastically reduces the coherence of the remaining answers. This scenario is obviously worsened by the removal of entire modalities, rather than individual questions, a not uncommon occurrence as shown in Table 2.1. This shows the quasi-

independent nature of the questions of medical questionnaires, even between different items of a same test.

2.4.3 Comparative Approaches

We compared our results with three comparative approaches. The simplest is *mean imputation*, in which each missing value is replaced with the mean of the corresponding variable. This is the most accurate solution when no internal representation (i.e. no information about the available data) is allowed to be used in the reconstruction process. PPCA [109] and IPCA [110] are the two others comparative approaches. As stated in Section 2.3.1, they are common PCA-based approaches to dimensionality reduction simultaneous with data imputation.

2.5 Experiments

We have three hypotheses to experimentally verify:

— H1: there is an optimal internal representation (IR) size to minimise error. That is, there is a degree of flexibility in terms of the networks' performance that diminishes the network's A_2 performance. This is expected for PCA in which, after a certain point, additional variables allow the IR to 'remember' missing values as the mean of their corresponding variable rather than impute them.

— H2: the network reconstructs data vectors more accurately when they are more complete, decreasing in performance as data becomes missing.

— H3: there is a benefit from learning from incomplete data points. Although complete data vectors are unarguably better, using case deletion to restrict the training dataset to only those data points reduce performance.

All p-values shown are post Bonferroni correction to account for multiple tests. All tests performed are paired t-tests. Prior to the experiments, all the data is normalized to the $[0, 1]$ range using each variable's theoretical maximum and minimum values. This equally weights variables regardless of discrepancies in their range.

H1: Optimal IR Size

This experiment consists of training and testing AE (and comparative PCA approaches) with IR sizes ranging from 1 to 10. In order to handle the low number of data-points, 20-fold cross-validation was performed to estimate the error. After splitting the dataset into 20 folds, one fold was iteratively selected as the testing dataset. The testing dataset was randomly corrupted with 10% chance of a modality being removed. This corruption process was repeated 40 times in order to have 40 differently corrupted versions of each testing dataset and total number of 800 data points for comparison. The datasets were saved to ensure that the same ones were used for evaluating each method, allowing for paired tests.

The remaining 19 folds were again split at each iteration with 80% as training data and 20% as validation data. The FCAE were re-initialised and retrained 8 times per iteration, and the one receiving the lowest validation A_2 error was selected to be evaluated on the testing dataset.

Dataset splitting was performed patient-wise, assigning all of the data from a patient into the same fold, implying that clinical records for one patient at two different times could not appear in both training and testing simultaneously. The validation dataset is corrupted in the same manner as the testing dataset to ensure that the validation loss represents both A_1 and A_2 testing error.

The A_1 and A_2 errors are shown in Figure 2.4. As expected, the A_1 error decreased monotonically with IR size for all methods ($p < 0.01$). As hypothesised, A_2 did show an optimal IR size for the PCA approaches with both methods monotonically decreasing until $IR = 4$ ($p < 0.01$) and monotonically increasing afterwards ($p < 0.01$ with the exception of PPCA between $IR = 7$ & 8). For FCAE, however, this consistent monotonic increase did not appear indicating that the learning process encouraged the AE to more actively impute that data. Our FCAE also performed better ($p < 0.01$) than both PCA-based methods for both A_1 and A_2 at all IR sizes indicating that there is some non-linearity in the underlying structure of the data.

H2: Predicting with Missing Data

This experiment was performed by training each method on all the available data and verifying that the testing error increases with an increasing corruption ratio. The methods were assigned an IR size of 4, corresponding to the optimal size determined in Section 2.5. The FCAE was trained with a corruption ratio of 10%, matching the lowest level of corruption performed. The results of this experiment are shown in Figure 2.5. As expected, the performance of all methods degraded as the corruption ratio increased ($p < 0.01$ between each consecutive corruption ratio, for every method), reflecting the decreasing amount of information available to the network to use in reconstruction.

The FCAE method outperformed both PCA approaches for both metrics at every corruption level ($p < 0.01$). FCAE's seem to be more sensitive to the amount of data missing compared to PCA, which may be the result of the discrepancy between the corruption ratio used during training and the one used in evaluation, significantly changing the training and testing distributions.

H3: Learning from Incomplete Data

This experiment consisted in training and testing FCAE and comparative PCA approaches, with IR size of 4, by discarding training and validation data samples that

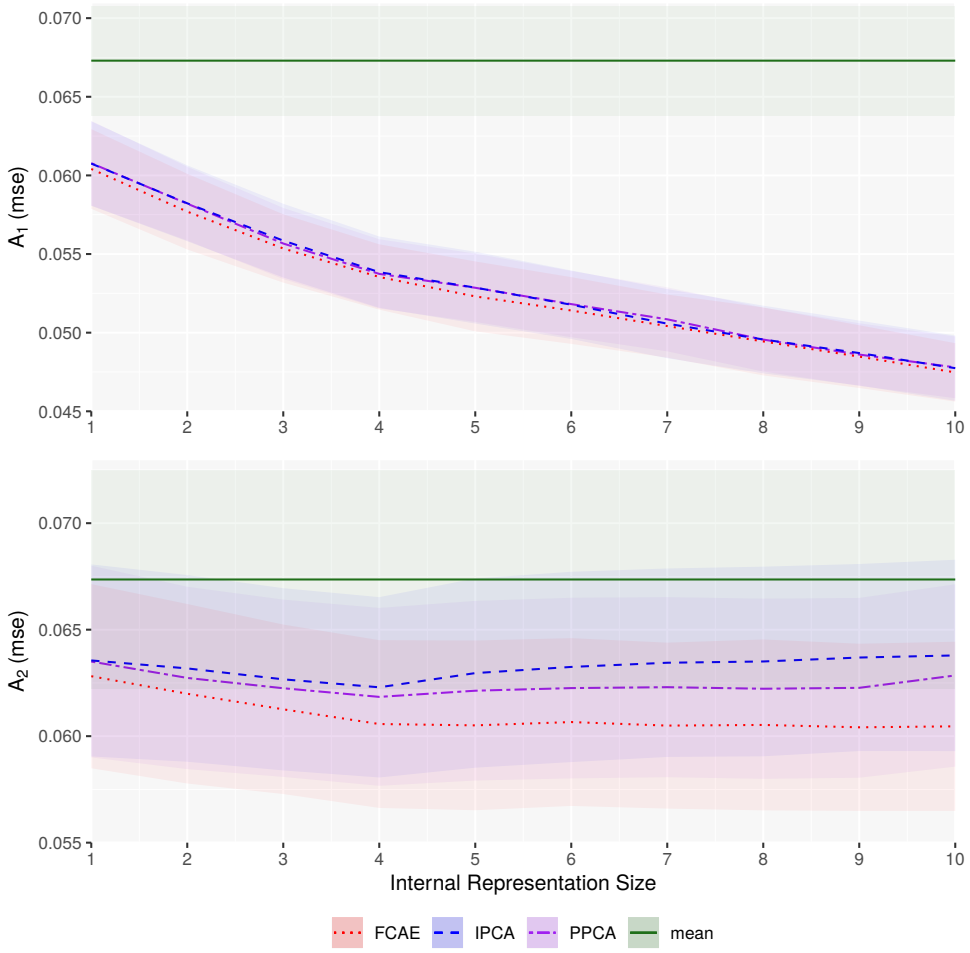


Figure 2.4 – A_1 (top) and A_2 (bottom) error for data reconstruction using FCAE (red) and comparative IPCA (yellow) and PPCA (blue) under varying IR size. The solid green line is the performance of mean data imputation with the dotted green lines representing the standard deviation thereof.

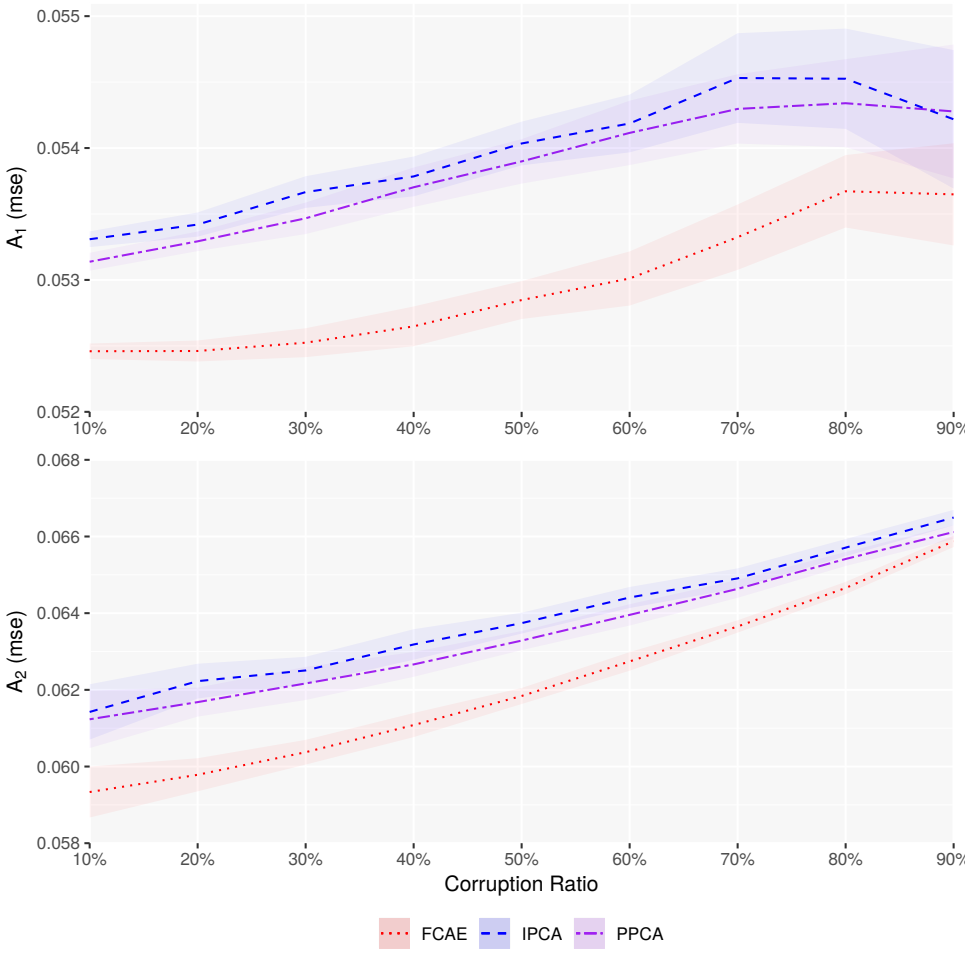


Figure 2.5 – A_1 (top) and A_2 (bottom) error for data reconstruction using FCAE (red) and comparative IPCA (yellow) and PPCA (blue) under varying levels of corruption.

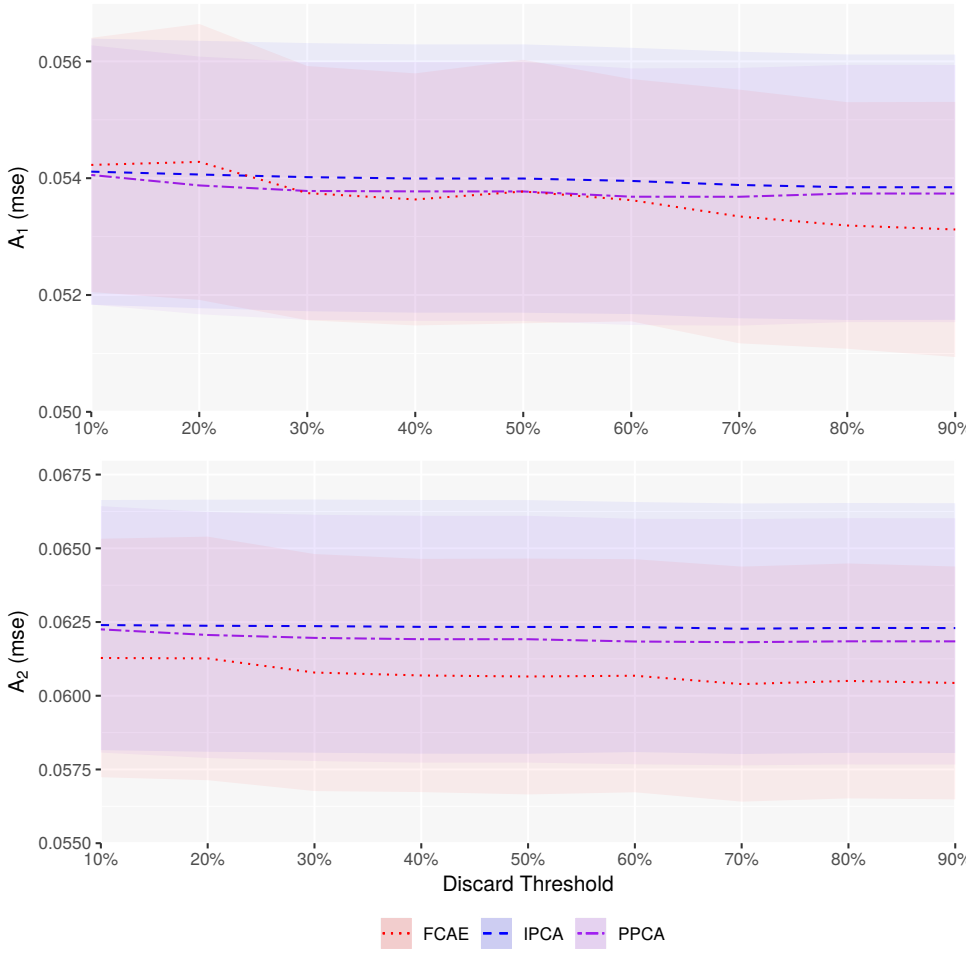


Figure 2.6 – A_1 (top) and A_2 (bottom) error for FCAE (red) and comparative IPCA (yellow) and PPCA (blue) methods with varying toss ratios.

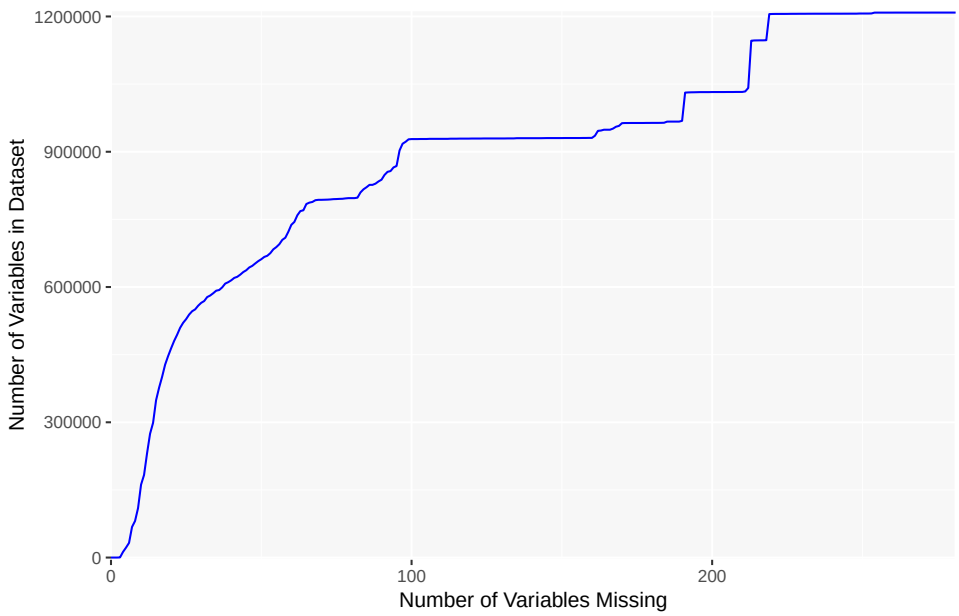


Figure 2.7 – Number of variables available when all rows with more than a certain number of missing variables (x-axis) are removed.

presents more than a fixed percentage of variables missing (which we will call the *discard threshold*). No testing data was removed. This was done in a cross-validation style similar to Section 2.5. The A_1 and A_2 results are shown in Figure 2.6.

Generally speaking, each method performed slightly better as more (although more incomplete) training data was provided. For all approaches, the improvement was very modest between consecutive discard thresholds with the exception of 10% and 20%. As shown in Figure 2.7, this is likely due to the larger number of added rows between those discard thresholds, compared to later thresholds. There was another swift increase between 60% and 80%, which can be seen in the improvements in the FCAE's A_1 and A_2 errors. This provides some evidence that training with incomplete data is beneficial, even if a sizeable portion of the dataset is largely incomplete.

It is interesting to note that the FCAE initially performed statistically significantly worse for A_1 than the comparative PCA approaches at the lowest discard thresholds (10% and 20%, $p < 0.01$), but significantly better at higher ones (70% to 90%, $p < 0.01$). This is likely because of the greater flexibility of the model, having more degrees-of-freedom and thus requiring more data to successfully fit.

2.5.1 Computation time

In order to quantify the computation time required for our method and the baselines, we ran a 20-fold cross-validation and measured the time required to compress and to uncompress the testing set. The results, in microsecond per sample, are displayed in Table 2.2, using an Intel Xeon E5-1620 v4 CPU at 3.50GHz.

Although our proposed autoencoder requires about 30 times more time to compress and uncompress a data sample, this would not be significant in a clinical setting as each method can be easily considered real-time.

Method	Comp. time (μs)	Uncomp time (μs)
PPCA	1.14	1.16
IPCA	1.05	1.27
FCAE	33.9	26.1

Table 2.2 – Average per sample compression and decompression time for our proposed autoencoder and the two baselines, with an Intel Xeon E5-1620 v4 CPU at 3.50GHz. All values are expressed in microseconds.

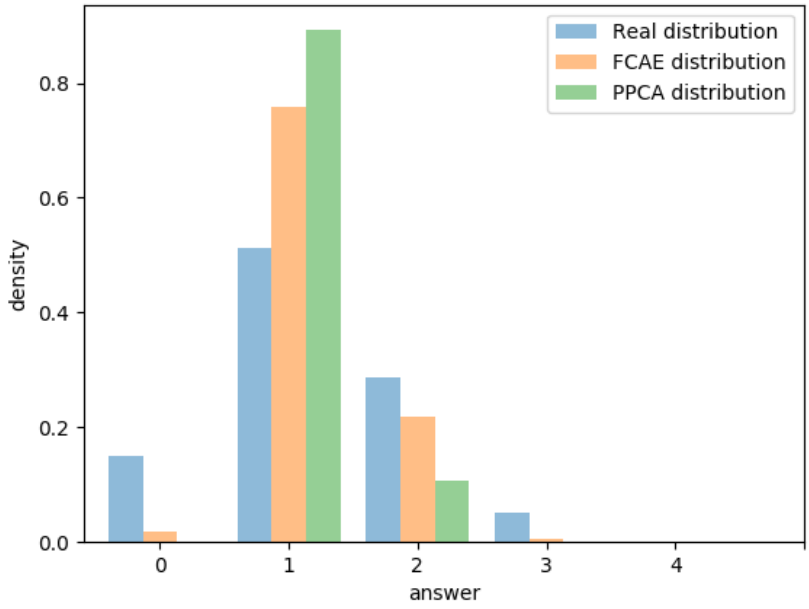


Figure 2.8 – True distribution of the first question of MDS-UPDRS part 3 (motor examination), compared with output distribution of FCAE and PPCA.

2.6 Discussion and Future Work

The first point of discussion is the relatively good performance of the two variants of PCA, indicating that a large amount of the variation can be addressed by a linear model equipped with simple missing-value handling. The difference between the two approaches to constructing linear models (PPCA and IPCA) was smaller than initially expected, with the more complex IPCA underperforming PPCA, illustrating that the linear approximation of the dataset's underlying structure is only that: an approximation.

The issue with using these linear models for reconstruction was expected: a sufficiently large internal representation allows PCA to *remember* the missing variables as their mean value rather than impute them. The deterioration of A_2 after a certain point in PCA is visible even at the relatively low IR sizes shown in Figure 2.4 (bottom). This trade-off is much less pronounced in the FCAE likely due to the application of modality masking noise which implies that the loss metric encodes both A_1 and A_2. This combination of losses means that the FCAE must take a more complex imputation strategy even at higher IR sizes.

We observed the distribution of the output prediction of each item, as seen in Figure 2.8. In this figure, we compare the output distribution of the first question of the part

3 (motor examination) of the MDS-UPDRS questionnaire as reconstructed using FCAE and PPCA on a test set against the true distribution as recorded in the database. 40 iterations have been performed, with a corruption ratio of 10% and IR size of 4. We can notice that there is an important loss of variance in the distributions, especially with PPCA reconstruction. We can explain that by the reconstruction algorithms being designed to minimize a variant of the mean squared error, thus biasing predictions values towards the mean of the distribution and penalising extrema. This behaviour is seen in both reconstruction methods although less distinctly in FCAE indicating that it has a higher capability of reconstructing these extreme values.

In future, we would like to extend our analysis of medical questionnaire imputation and compression for a particular downstream analysis goal, such as the classification of different Parkinsonian patient groups or the stratification of the disorder. At the moment, diagnosis and stratification are done by a neurologist using some of the data provided by the medical questionnaire, indicating the utility of making the analysis thereof more objective and robust to heterogeneous, incomplete data.

One fact that complicates this study is the relatively low number of data points, especially complete data points, given the relatively high dimensionality and lack of simplifying structure in each point. These issues are further complicated by the majority of tests in the PPMI database used having a middling number of ordinal values (most variables took on values in $\{0..5\}$) rather than simpler categorical or continuous values which thus limit their utility as input.

Ordinal and categorical data handling

It is important to note that, with this database, all the variables were numerical or ordinal. Thus, we did not test the framework with categorical variables, which can be present in other medical questionnaires. We would suggest the use of one-hot encoding for this type of variable in addition to a custom output layer capable of handling different loss functions for the different variable types.

Imputing ordinal variables is problematic. Although ordered, the leap between two successive categories is not consistent, nor quantified. Thus, an error in imputation for an ordinal variable can be more or less significant in certain parts of the answers range. In medical questionnaires, some questions suffer greatly from this problem, as a ‘0’ could mean complete absence of a symptom, and ‘1’ to ‘4’ could quantify its severity. To this extent, the difference between ‘0’ and ‘1’ has more impact on clinical interpretation than

the difference between '1' and '2'. Neither our method, nor the baselines, straightforwardly address the specificities of ordinal variables, but instead treat them as continuous, a common approach in the literature [114]. One way of tackling this problem with our framework would be to use a loss function tailored for ordinal variables, such as the weighted kappa loss proposed by De la Torre *et al.* [120]. Ideally, the weights of each consecutive error of each ordinal variable should be defined by clinicians according to its impact on clinical interpretation, thus leading to a greater practical applicability.

An alternative approach would be to use an adversarial loss that could, in theory, learn what continuous values correspond with possible levels for each ordinary variable and well as what combination of values are realistic [99]. Such an approach would require a significant database consisting of full rows, i.e. data that can be used as true fully sampled data for training the discriminator. However, this is not representative of the PPMI database as described in Section 2.4.2.

2.7 Conclusions

This Chapter presents an autoencoder specifically designed for data imputation and compression of medical questionnaires, in which entire modalities may be missing. This represents some of the initial steps into performing deep learning on specific medical tasks that are challenging due to the dataset size compared to the high number of features, the data types provided and the complexity of the learning problem. We have shown that significant and consistent improvements can be made over linear methods, especially when a lower number of features is missing. We have also shown that there is an interest in learning on incomplete data vector, and that the imputation performances of our framework is not negatively impacted by a high internal representation size, making the downstream choice of the compression rate easier.

Acknowledgments

Data used in the preparation of this Chapter were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (www.ppmi-info.org/data). For up-to-date information on the study, visit www.ppmi-info.org. PPMI - a public-private partnership - is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including AbbVie, Allergan, Avid Radiopharmaceuticals, Biogen, BioLegend,

Bristol-Myers Squibb, Celgene, Denali Therapeutics, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Prevail, Roche, Sanofi-Genzyme, Servier, Takeda, Teva, UCB, Verily, Voyager Therapeutics and Golub Capital.

Chapter conclusion

In this chapter, we developed and validated a method to compress clinical testing data that is robust to missing values. This method is a crucial element of our patient-specific DBS outcome prediction pipeline, as it allows for a maximum number of clinical tests to be considered, and therefore a maximum of the patient's clinical functions, in a prospective manner. Indeed, our method can project a patient's clinical state into a low-dimensional, continuous latent space, allowing for downstream application to use this more readily usable representation of the data without suffering from the incompleteness and high dimensionality of the original, raw input space.

Chapter 3

ParDi: Parkinson's disease stage classification using deformation of the striatum

As well as medical exams and questionnaires, clinicians also review medical imaging scans to better understand the patient's disease and determine a course of treatment. For example, large brain alterations are a counter indication for DBS, making medical imaging an important modality in input of our patient-specific CDSS for the prediction of DBS clinical outcomes.

The number of cortical and subcortical structures of interest are numerous, as well as the number of possible medical imaging scans. Among the different scans, we retained the T1-weighted MRI scan because it is used universally for neurological disorders and provides good visualization of our structures of interest, the bilateral putamen and caudate nuclei. These structures play a large role in the patient's cognition and is believed to change shape as the patient's disease progresses.

In contrast to most of the literature which tends to simplify striatal alteration to one score (volume or thickness), or to do a voxel-wise analysis, we will, in this chapter, consider the entire shape displacement field of each structure simultaneously. We will quantify their relevance as a staging biomarker in order to determine if striatal shape information is an appropriate input for our DBS clinical outcome prediction system.

This chapter has been published in NeuroImage: Clinical, 2020, vol. 27 [1].

1. Peralta, M., Baxter, J.S.H., Khan, A. R., Haegelen, C., and Jannin, P. (2020). Striatal shape alteration as a staging biomarker for Parkinson's Disease. NeuroImage: Clinical, 27.

3.1 Abstract

Parkinson's Disease provokes alterations of subcortical deep gray matter, leading to subtle changes in the shape of several subcortical structures even before the manifestation of motor and non-motor clinical symptoms. We used an automated registration and segmentation pipeline to measure this structural alteration in one early and one advanced Parkinson's Disease (PD) cohorts, one prodromal stage cohort and one healthy control cohort. These structural alterations are then passed to a machine learning pipeline to classify these populations. Our workflow is able to distinguish different stages of PD based solely on shape analysis of the bilateral caudate nucleus and putamen, with balanced accuracies in the range of 59% to 85%. Furthermore, we compared the significance of each of these subcortical structure, compared the performances of different classifiers on this task, thus quantifying the informativeness of striatal shape alteration as a staging bio-marker for PD.

3.2 Introduction

Some non-motor clinical symptoms may be detectable before the appearance of the more distinctive motor symptoms [121]. This prodromal stage of PD is crucial for medical treatment but difficult to diagnose. There are a variety of treatments for PD, including pharmacological treatment such as the dopamine precursor levodopa or interventional treatments such as DBS, which are proposed to obstruct disease progression and enhance the patient's quality of life. The choice of treatment remains under debate, and the high heterogeneity of the disease [122] renders the space of treatment options highly variable and patient-specific [123] [124]. Finally, there are unanswered questions regarding the long term effects of these treatments and the evolution of the disease [125]. For all of these reasons, PD is unanimously considered in pressing need of biomarkers both for improved diagnosis and treatment monitoring [126].

In terms of the underlying neuroanatomy of Parkinson's disease symptomology, promising results have been found in the bilateral putamen and caudate (the two major components of the dorsal striatum). It is well known that these structures are central to the progression of PD, which greatly alters their behavior. The lack of dopamine in the putamen as a result of PD is considered the direct cause of motor dysfunction, while the lack of dopamine in the caudate is more related to alterations in cognitive function [127]. Griffiths *et al.* [128] reported an alteration of the density of some neurotransmitter receptors in both of these structures in post mortem brains of PD subjects. Kish *et al.* [127] observed a large loss of dopamine in both putamen and caudate nucleus, with reductions of over 99% in caudal portions of the putamen.

In clinic, the diagnosis of PD, staging of the disease, and selection of the treatment are made primarily using clinical biomarkers such as UPDRS scores or Hoehn and Yahr scale. Although brain alteration is prior to any clinical symptom, brain morphometry via MRI is considered a marginal source of information in clinic PD [129], [130]. With increasing availability and resolution of MRI, the research community is moving towards finding reliable imaging biomarkers for the diagnosis and monitoring of PD.

Subcortical morphological biomarkers

Shape alterations in subcortical structures aside from volume or thickness reduction has long been identified as a potential area of analysis. Voxel Based Morphometry (VBM) has been for a long time the favoured method due to its fine granularity, although with de-

fined limitations [131], [132]. Most modern methods fall into the category of Surface-based Morphometry (SBM) which is a two-stage process. The first stage involves the segmentation and registration of subcortical structures from the MRI into a template space, and the second stage involves the representation and analysis of the segmented surface's shape with respect to that of the population. While there are a variety of segmentation methods in the first stage, the second almost unanimously extracts vertex-wise boundary displacements on a template mesh, representing each surface as a large, but constant-sized, vector. Among the most used, the Bayesian Appearance Model (BAM) is now a native part of the FSL library using the FIRST implementation. This method proposes to add a Bayesian framework to the Active Appearance Model (AAM), which incorporates intensity information on top of the shape deformation model. The authors claims that this Bayesian framework allows to capture more subtle shape deformation than the other methods, even when the training data amount is low. Other widely used methods for shape analysis are single-atlas segmentation or multi-atlas label fusion, followed by Large Deformation Diffeomorphic Metric Mapping (LDDMM) [133] for registration.

Amongst the techniques explored to analyze the brain morphometric variations, shape analysis has proved itself more reliable and consistent than volume analysis, thickness analysis, and voxel-based morphometry [129], [134]–[139] and is considered as a promising source of insight into various neurological and psychiatric disorders.

Subcortical shape analysis in Parkinson's Disease

Several studies have tried to evaluate the significance of subcortical shape displacements as a bio-marker for PD, leading to often contradictory results with the literature showing the impact of cohort and methodology on the relevance of subcortical structures as diagnosis biomarkers. Garg *et al.* [140] show that the morphology of the bilateral caudate nucleus as well as the putamen is discriminant in PD. Nemmi *et al.* [141] found significant results only in the left caudate and putamen, whereas Owens *et al.* [142] didn't find any correlation in shape information outside of gross volumetric differences.

Some tried to use subcortical shape displacement for other purposes than classifying PD patients from healthy controls. Mak *et al.* [143] and Foo *et al.* [144] attempted to correlate subcortical shape displacement with Mild Control Impairment (MCI) in PD, the first not showing significant results while the second found significant information only in the left caudate. Both Nemmi *et al.* [141] and Garg *et al.* [140] tried to correlate subcortical shape displacement with disease severity (using left, right and global UPDRS

and disease duration for the first, and global UPDRS only for the second). The only significant result obtained, on these two studies, was between the right UPDRS and the left putamen shape [141]. Owens *et al.* [142] failed to show discriminating shape features in the putamen and caudate nucleus to classify different stages of PD, getting only significant results through volume analysis. Ultimately, the literature shows that correlation between specific clinical symptoms and subcortical shape displacements in PD remains a difficult task.

Contributions

To the best of our knowledge, no cross-validated studies has been done to quantify the predictive power of subcortical shape displacements in PD. Moreover, little work has been done to compare the relevance of classifiers and to compare the informativeness of the structures. Finally, most of the studies focus on diagnosing PD against Healthy Control (HC), without differentiating intermediate stages, and the need to assess the relevance of morphometric biomarkers to diagnose PD prodromal phase has been pointed out [140].

Our proposed method is the first fully-automated, cross-validated pipeline to classify different stages of PD. This data-driven pipeline uses the state of the art of machine learning and data analysis to quantify the relevance of shape displacements of putamen and caudate nucleus as diagnostic and staging biomarkers, benchmarking different classifiers and structures. We have extended clinical knowledge by investigating subcortical shape displacements and determining those which are relevant for diagnosing prodromal stages of PD, therefore responds to the necessity raised by Peran *et al.* [129] to extend the morphometric biomarkers research on prodromal stage of the disease.

We also shown that the MDS-UPDRS3 score of PD patients can be predicted from subcortical shape displacements with our method, showing weak, yet significant (at $p < 0.001$) results.

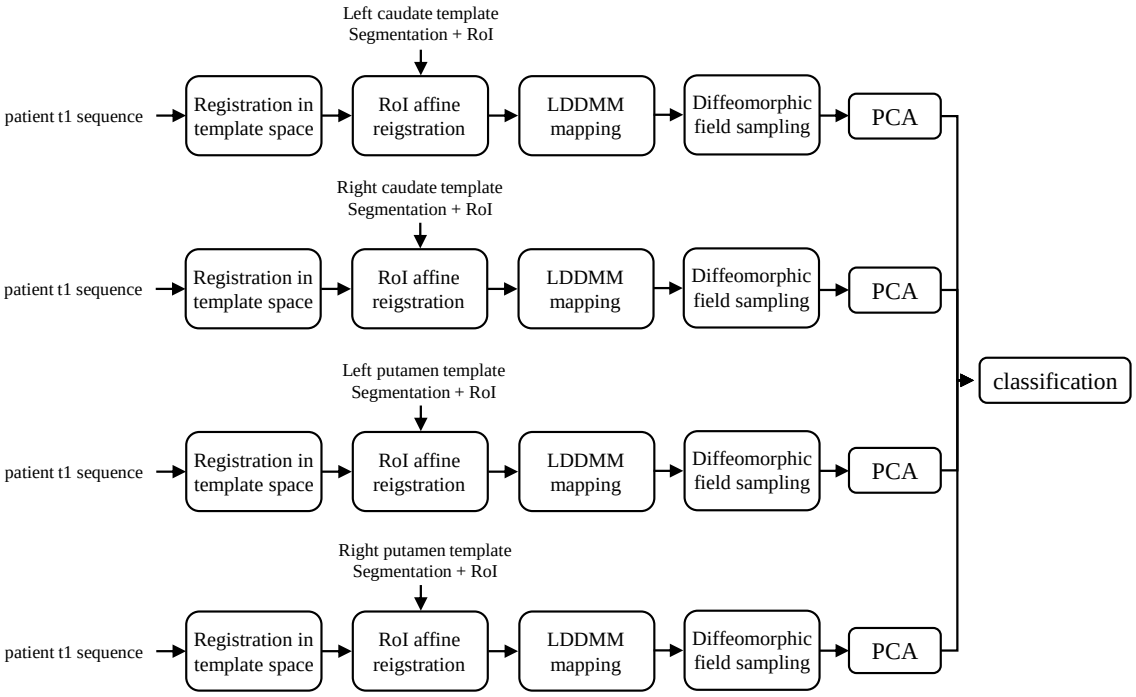


Figure 3.1 – Pipeline proposed and tested in this study.

3.3 Materials and Methods

3.3.1 Proposed Method

The pipeline proposed is composed of three steps, as presented in Figure 3.1, which include the extraction of displacement vectors, their compression, and finally the classification of the patient based on this information.

The first step is to extract the displacement fields of the four striatal structures from T1-weighted MRI sequences, compared to a fixed atlas template described in Section 3.3.3 using the method presented in Section 3.3.1. The output of this process is a vector of vertex-wise signed displacement magnitudes, with positive values indicating displacement outwards from the template surface and negative indicating an inwards displacement. We chose not to analyze the full displacement vectors themselves, as we hypothesize that taking into account the direction of the vectors would not add substantial information, while increasing the vector size by a factor of three.

As the vector of displacements magnitudes is high dimensional (6362, 6391, 6490, and 6510 elements for left and right caudate and left and right putamen respectively), a compression step was performed prior to the final classification step. Each structure has

been compressed through PCA. The number of principal component kept has been chosen through HPO, as explained in Section 3.3.6.

Finally, a classification is performed using the concatenation of the compressed displacements magnitudes vectors as input. These methods are explained in more detail in the subsequent sections.

Image processing pipeline

We used a single-atlas based segmentation followed by LDDMM registration and label propagation to automatically extract subcortical shape displacements. More details about this image processing pipeline can be found in [145], and the full scripts have been made available by the authors (`https://github.com/khanlab/diffparc-sumo`).

The original version of this method was proposed by Khan *et al.* [146] and shown a high degree of robustness for the segmentation of the caudate nucleus and putamen with a Dice overlap measure of 81% and 83%. As the reliability of this method has already been proven many times [140], [145]–[149], we did not perform any extensive segmentation quality check at this stage.

To summarize, this pipeline first registered the T1-weighted MRI scan into MNI152NLin2009cAsym space. Then, a Region of Interest (RoI) around the desired structure of interest on T1-weighted MRI scans was coarsely aligned with a template segmentation using an initial affine registration step. Finally, this mapping is deformably refined using LDDMM. The output of this process is a smooth diffeomorphic vector field, which can be sampled at the surface vertex locations in order to create the surface displacement vector. The magnitude of each vertex displacement is then saved for each surface.

Compression

We compressed displacements vectors through PCA. By keeping only the first few principal components which are orthogonal and therefore decorrelated, the data is compressed while keeping most of the information. On top of that, it often eliminates noise in the data, as noise (assuming it is of lower variation than the data and is, independent of it) is often relegated to the least-significant principal components. This triple advantage of compression, decorrelation, and denoising makes PCA useful for downstream statistical or machine learning analysis, especially when the original dimensionality of the data is high compared to the number of available samples.

Classification

We used and compared four different classifiers in this study: SVM with linear and radial basis kernels, RF and EL through stacking classifiers. Stacking classifiers combines the strengths of different classifiers. The meta-classifier may be trained solely on the predictions of the base classifiers or to rather extend the original pool of features with them. Although the former is simpler and has much lower parameterization, the latter allows the meta-classifier to use the underlying data to determine which classifiers are likely to be more accurate for any particular data point. In this study, we chose the later option in order to give the meta-classifiers the theoretical ability to dynamically adapt the weight of base-classifier, depending on the inputs. We used logistic regression as a meta-classifier.

3.3.2 Data

Cohort description

Data for this article come from two databases. The first database consists of PD patients from the cohort of DBS patients recruited at Rennes University Hospital (referred to as 'DBS PD'). These patients are all candidates for DBS, indicating that they experience sufficiently advanced motor symptoms to warrant such an intervention and are thus in an advanced phase of the disease. All patients of this cohort have been informed and gave their consent to be included in this study, which have been approved the institution's ethics committee (trial code '35RC19_4001_PSCP'). The second database is derived from the Parkinson's Progression Markers Initiative (PPMI), which is a program sponsored by the Michael J. Fox Foundation for Parkinson's Research. It is an observational clinical study which tracks cohorts of subjects with different forms of Parkinson disease for up to 8 years, with the goal of identifying biomarkers of disease progression using MR imaging, biologic sampling as well as clinical and behavioural assessments.

The PPMI database is used to define three separate cohorts. The first consists of HC subjects without any neurologic disorder and who do not have an immediate relative with PD. The second is a prodromal stage cohort. Prodromal PD patients are at risk of developing PD and display some characteristic clinical symptoms but not the more diagnostic motor symptoms [121]. The third is the early PD cohort, composed of subjects with a diagnosis of PD for two years or less, and not taking PD medications.

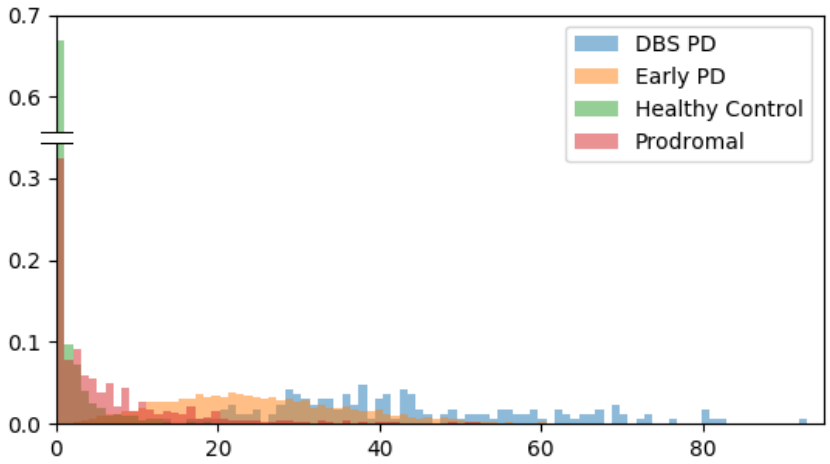


Figure 3.2 – UPDRS-3 normalized distribution of the cohorts used in this study.

Cohort	F/M (total)	Age (range)	Mean UPDRS (range)
HC	64/113 (177)	67.8 ±11.2 (39-90)	1.73 ±3.35 (0-34)
Prodromal	9/32 (41)	74.9 ±6.93 (56-91)	6.73 ±8.60 (0-52)
Early PD	127/241 (368)	68.4 ±9.73 (40-98)	25.4 ±11.7 (0-90)
DBS PD	76/104 (180)	65.1 ±9.48 (26-85)	43.4 ±16.1 (14-92)

Table 3.1 – Statistics of the cohorts used in this study.

Figure 3.2 shows the normalized distribution of UPDRS3 score, the main motor assessment for PD, for the different cohorts. The UPDRS3 score used for the DBS PD cohort has been converted to the MDS-UPDR3 score used in PPMI with the equation proposed by Hentz *et al.* [150]. The three PPMI cohorts follow well-defined exponential or Gaussian distributions, whereas the Rennes cohort is more heterogeneous. This heterogeneity is expected as DBS is now sometimes proposed quite early during the disease at the Rennes University Hospital. More information about the inclusion criteria of the PPMI cohorts can be found on the PPMI study protocol, following the link `https://www.ppmi-info.org/study-design/research-documents-and-sops/`.

Table 3.1 shows the UPDRS mean and standard deviation of each cohort. From this, one can observe a natural progression of the disease in terms of motor symptoms: first healthy, then prodromal stage, then early-stage Parkinson and finally late-stage Parkinson. One can also note that the standard deviation of each cohorts is quite high, indicating that this information alone is not optimal and underscoring the importance of finding other biomarkers.

Image acquisition

All patients from the Rennes database had one preoperative 3T T1-weighted MRI scan (1mm x 1mm x 1mm, Philips Medical Systems). All sequences were acquired prior to DBS electrode placement.

For the PPMI database, all T1-weighted sequences (e.g. MPRAGE or SPGR) were required to have a total scan time between 20 and 30 min and to have a slice thickness of 1.5 mm or less with no interslice gap.

3.3.3 Atlas

Probabilistic segmentation of striatal structures was obtained with the MNI PD25 atlas [151]. This atlas was built by averaging 3T MRI scans (T1w (FLASH and MPRAGE), T2*w, T1-T2* fusion, phase, and an R2* map) of 25 PD patients, making it an atlas of choice for studying PD patients in the MNI152NLin2009cAsym space. Eight subcortical structures have been segmented in this atlas, including the caudate nucleus and the putamen. This atlas is freely available through the link `http://nist.mni.mcgill.ca/?p=1209`.

As noted in Section 3.2, the striatum has been of particular interest in morphological image analysis for PD. Thus, the segmentation of the left and right caudate and putamen were extracted for use in our study.

3.3.4 Accuracy and loss metrics

As cohort sizes are uneven, we used Balanced accuracy (BACC) (eq. 3.1) as a classification performance and comparison metric. This balanced accuracy is necessary to prevent systemic bias towards the larger cohort during the classification step. For compression, we used the reconstruction mean standard error as a compression performance and comparison metric. Each classifier was trained with class weighting, giving a weight to training sample inversely proportional to the representation of the belonging class in the training set.

$$BACC = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right) \tag{3.1}$$

3.3.5 Training and validation

For each test, we used a stratified 10-fold CV. Each model has been trained 10 times separately, using one fold as a validation data and the remaining nine as training data. Each model has been trained and evaluated with the same folds, removing this as a potential source of variation. Additionally, each fold contains approximately the same number of samples from each cohort. In the event that a patient has multiple MRI acquisitions, all the acquisitions are assigned to the same fold, in order not to ensure the folds are independent of each other and there is no training-evaluation set corruption. To address the issue of class-balance in training, each classifier has been trained weighting samples by the inverse of the class size.

3.3.6 Hyper-parameter optimization

The hyper-parameters of each classifier have been optimized for each CV fold through Bayesian optimization using a Gaussian process as a surrogate model with expected improvement as the criterion. The number of points tested was the number of hyper-parameters to optimize squared plus one.

3.3.7 Statistical analysis

The resulting BACC's were analyzed statistically using multi-factorial ANOVA in order to estimate the effect size and possible contributions of different classifiers and combinations of structures, our literature review suggesting these as potential sources of variability.

3.3.8 Software environment

The Scikit-learn implementation of PCA as well as each classifier was used, and the Python code-base used to perform the different experiments is available at `https://github.com/m-prl/ParDi`. Statistical analysis was performed using IBM SPSS Statistics.

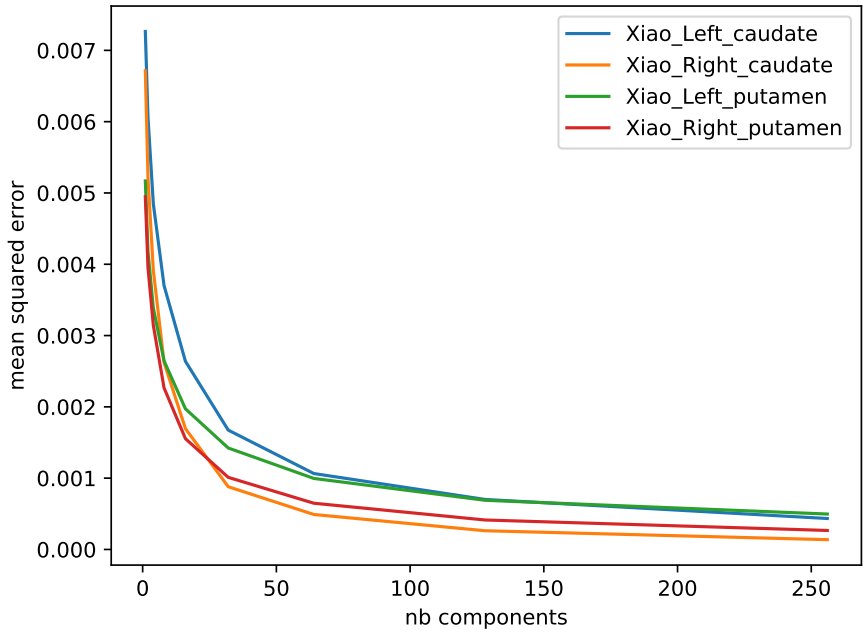


Figure 3.3 – Reconstruction mean squared error of PCA compression on test set for left caudate nucleus (blue), right caudate nucleus (orange), left putamen (green) and right putamen (red), with various number of components kept.

3.4 Results

3.4.1 Compression performance

Figure 3.3 presents the PCA compression mean squared errors for all four structures, and with different numbers of components kept.

3.4.2 Classification results

We performed a total of 1200 binary classifications (10 folds times 6 cohort-pairs times 5 combinations of structures times 4 different classification algorithms) each involving an independent hyper-parameter optimization process. We reported the best balanced accuracy configuration of 10-fold classification. Table 3.2 presents the multi-factorial ANOVA analysis of the classification results, showing that every variable contributed to the results with a high level of significance.

As expected, the Problem (pair of cohorts being distinguished) factor had a more defined effect over the Structs (the combination of structures used for classification)

Source	T.III SS	df	Mean Sqr	F	Sig.	ηp^2
Corr. Model	7.31[a]	12	0.61	260.7	0.00	0.81
Problem	6.80	5	1.4	832.5	0.00	0.78
Structs	0.31	4	0.08	47.12	0.00	0.14
Structs*Problem	0.52	20	0.03	15.93	0.00	0.22
Algo	0.21	3	0.07	42.11	0.00	0.10
Fold	0.36	9	0.04	24.66	0.00	0.16
Error	1.89	1158	0.02			
Corr. Total	10.08	1199				

Table 3.2 – Multivariate ANOVA test of the BACC across methods organized by problem, combination of structures used, and classification algorithm.
a. R Squared = 0.812 (Adjusted R Squared = 0.806)

Problem	BACC	Sens.	Spec.	F1
Early vs HC	95%	99%	82%	97%
Early vs Pro.	82%	98%	27%	86%
Pro. vs HC	64%	69%	77%	46%
DBS vs Pro.	94%	85%	98%	90%
DBS vs Early	72%	11%	98%	19%
DBS vs HC	99%	85%	100%	92%

Table 3.3 – Results of different binary problems using MDS-UPDRS3 score as input and a naive Bayes classifier.

and Algo (classification algorithm) factors. We can also note the significance of the Structure∗Problem factor, indicating that the structures of interest vary through the stage of the disease.

In order to further investigate the effects of the structures and classifiers used, follow-up analysis using Tukey's Honestly Significant Difference (HSD) test was performed. This test partitions the values for each factor (e.g. Structure, Classifier, etc...) into a series of clusters that are statistically significantly different from each other. This allows for us to infer which structures or classifiers have improved performance given that significance under the ANOVA test.

Problem	BACC	Sens.	Spec.	F1
Early vs HC	59%	77%	38%	67%
Early vs Pro.	64%	93%	23%	88%
Pro. vs HC	69%	42%	88%	49%
DBS vs Pro.	78%	89%	62%	86%
DBS vs Early	80%	64%	91%	68%
DBS vs HC	85%	87%	87%	83%

Table 3.4 – Results of different binary problems, with Ensemble Learning as a classifier, and all the structures in input. The first class of the problem is considered the positive class.

Classification performance between problems

In order to develop a reference of the relative difficulty between the problems, a preliminary classification using only the clinical symptomatology was performed. Table 3.3 presents the classification metrics obtained on the training base for the 6 different binary classification problems, based solely on the MDS-UPDRS3 using a naive Bayes classifier. It is to note that the dataset is heavily class unbalanced between the DBS cohort and the others, as in the PPMI program the MDS-UPDRS3 is assessed at multiple visits, thus explaining the low f1 scores and sensitivity of the last three rows of the Table. The accuracy is expected to be higher using clinical symptomatology than via the use of morphological information as clinical symptomatology is currently relied upon for the diagnosis and staging of PD [121]. Imaging, let alone morphological analysis, is not always used in the diagnosis of PD. This is particularly relevant for defining Early and Late stage PD in which the motor symptoms are a defining characteristic of the disorder [121]. Thus, one would expect nearly perfect accuracy between these two stages and Prodromal or HC based solely on motor symptomatology, although this gives little information about the underlying etiology of Parkinson's disease, only the clinical workflow used to diagnose it.

Table 3.4 presents different performance metrics for all the six problems using all structures as input and Ensemble Learning as the highest performing classifier. Further statistical analysis (i.e. Tukey's HSD test) could not be performed as the cohort size plays a preponderant role in the analysis of these results. It shows that logically, the DBS cohort was the easiest to classify, as it is composed of patients with the most advanced state of the disease. The three other cohorts were more difficult to distinguish between themselves.

An interesting result was that the use of morphological information for classifying between Prodromal stage PD and healthy controls appears to outperform the use of motor

clinical symptomatology. This is likely as the motor symptomatology does not capture the other symptoms that are more indicative of Prodromal stage PD which may be related to morphological changes detectable from imaging.

It however is unexpected that the HC versus Early PD problem was more difficult than the Prodromal versus Early PD, as the prodromal stage is widely considered anterior than the early stage, in the course of the disease. Yet, prodromal stage is considered as very heterogeneous in terms of symptomatology [121].

Classifier comparison

Table 3.5 shows the BACC's across the different classification algorithms as well as the results of a Tukey's HSD test. It shows three clusters of classifiers with algorithms in different clusters having statistically significant differences in performance. The first cluster is composed of RF alone indicating that it has significantly worse performance of all the methods investigated. The second cluster is composed of SVM with a linear kernel indicating that although it significantly outperformed RF, it still underperformed SVM with a non-linear kernel and ensemble learning. The final, highest-performing cluster is composed of the SVM with radial basis kernels and Ensemble Learning (EL) indicating that these two methods are roughly equivalent in performance, neither outperforming the other significantly.

The fact that EL does not have significant better results than SVMr can be explained by the fact that it often relied solely on the prediction of SVMr, which is the highest accuracy base classifier. Indeed, the three base classifiers mostly agree on the classification outcomes. This lack of variety in the predictions between base classifiers limits the interest of stacking these classifiers together. Nevertheless, stacking classifiers can be considered as a safer choice in general as it theoretically and experimentally does not under-perform any base classifier. Overall, our results indicate that although different classification algorithms delivered statistically significantly different performance, this effect is marginal.

Structure comparison

Table 3.6 shows Tukey's HSD test results to compare the informativeness of the four different structures studied identifying three clusters. The first and lowest performing cluster included the left and right caudate as well as the left putamen, implying that these structures provide an equivalent amount of information to the classification algorithm. The second cluster contained the right putamen, which was statistically significantly more

Classifier	N	Cluster		
		1	2	3
RF	300	68.0%		
SVMl	300		69.3%	
SVMr	300			70.9%
EL	300			71.3%
Within-group Sig.		1	1	0.674

Table 3.5 – Tukey's HSD test to compare classifier performances. Mean BACC for each classifier is also displayed. Alpha = 0.05.

Structures	N	Cluster		
		1	2	3
Left caudate	240	68.3%		
Left putamen	240	68.7%		
Rigth caudate	240	69.3%		
Right putamen	240		70.6%	
All structures	240			72.7%
Within-group Sig.		0.224	1.000	1.000

Table 3.6 – Tukey's HSD test to compare structures' performances. Mean BACC for each structure combination is also displayed. Alpha = 0.05.

informative than the other individual structures. The final and highest performing cluster reflects when the classifier had information regarding all structures simultaneously.

It is unsurprising that taking all four structures together to make predictions is statistically significantly better than taking any individual one (Table 3.6). This demonstrates that there exists some non-overlapping information across structures, i.e. that the effect of PD cannot be isolated solely to a single structure, and that striatal structure deterioration is not uniform between all structures. Subcortical structure deterioration in PD is likely to be specific to each structure, with these specificities being differently informative. It was unexpected, however, that right structures seem to be more informative overall than the left ones which is investigated in the following section.

3.4.3 Laterality significance

As shown in Section 3.4.2, right structures appeared to be more informative than the left ones when considered in isolation. This asymmetry could be due to two sources: a consistent right-left difference in the algorithm or underlying disease progression, or

a contralateral-ipsilateral difference which is exposed from a left-right bias in the underlying population. A possible technical reason for the first source could be the higher compressibility of right structures, shown in Section 3.4.1. As fewer principal components are needed to embed more information about the shape displacements for right structures, it is possible that classifiers favour these structures more than the left ones.

The second source is partially confirmed by the slight over representation of right-side PD in our DBS PD cohort (81 left side start against 86 right side start). In this cohort, we could extract information regarding the symptom progression of PD, determining whether or not the symptoms began on the right or left side. Indeed, if the disease starts evolving earlier on one side of the brain, the subcortical structures of this side may be in a more advanced phase of deterioration, and thus are more informative to diagnosis the disease.

To verify this hypothesis, we ran a new experiment to compare the significance of left against right structures for both left sided PD patients and right sided PD patients. We ran a 10-fold CV, using Ensemble Learning as a classifier, as it is the most performant, as shown in Table 3.5. Two problems have been tested, HC versus Left sided PD and HC versus Right sided PD. For each problem, we reported the BACC of the 10-fold CV, using in one case the left caudate nucleus and the left putamen, and in the other case the right caudate nucleus and right putamen. For each new problem, a paired-sample t-test has been performed to assess if the side ipsilateral to disease progression is giving statistically better accuracy results than the contralateral side. Results are reported in Table 3.7. We can see that, for left-sided PD patients, the left structures were consistently more informative than the right ones ($p = 0.009$). For right-sided PD patients, right structures were consistently more informative than left ones ($p < 0.001$).

This experiment confirms the hypothesis that the starting side of PD is an important source of diagnostic information, and should be taken into account when interpreting medical images. This suggests that the left-right difference observed in Table 3.6 is a result of a difference in population size of left-sided and right-sided PD in which there is a contra- vs. ipsi-lateral difference in terms of the information content of each structure. However, more longitudinal studies would be necessarily to determine if this difference persists in the Early and Prodromal cohorts in which this laterality may not have yet manifested.

Struct. side	Left struct.	Right struct.	p-value
HC vs Left PD	81.8% ± 0.5	78.5% ± 3.4	0.009
HC vs Right PD	76.5% ± 1.2	87.9% ± 2.1	<0.001

Table 3.7 – BACC for Ensemble Learning between HC and DBS cohorts, the latter having laterality information regarding PD progression.

Figure 3.4 – 10-fold CV MDS-UPDRS3 prediction with proposed method, for cohorts DBS PD, Early PD and Prodromal. Red curve is the linear regression line of the predictions.

Method	MSE	MAE	R
Proposed method	224.4 ± 380.4	11.64 ± 9.434	0.215
Mean baseline	234.9 ± 433.1	11.77 ± 9.816	-0.160

Table 3.8 – Statistics for 10-fold CV MDS-UPDRS3 prediction with our method and a mean-prediction baseline, for cohorts DBS PD, Early PD and Prodromal.

3.4.4 MDS-UPDRS3 prediction

In order to quantify the predictive power of striatal shape morphology for the motor symptomatology of the patients, we used an analogous pipeline as in the previous experiment, with RF as a regressor instead of a classifier, to predict the MDS-UPDRS3 score of the patients of all cohorts, besides HC. We decided to remove the HC cohort from this experiment, as the task of predicting their MDS-UPDRS3 is not relevant, as it is concentrated at zero, thus biasing the model's prediction towards zero. Figure 3.4 presents the results of a 10-fold CV, after having ran an HPO. We compared the performance of our system with a baseline predicting always the cohorts' mean MDS-UPDRS3 score. The performance of both systems are shown on Table 3.8. The correlation coefficient R of the regression is equal to 0.215, which can be considered as weak, yet significant ($p < 0.001$). This may indicate that the investigate subcortical shape alterations are co-caused along with motor symptomatology, rather than being in direct causation.

3.5 Discussion

When comparing the baseline classification results of Table 3.3 exclusively based on motor symptomatology, and the results of Table 3.4 obtained with our system, we can observe that, in 4 cases out of 6, the results are better for the baseline. We can explain that by the fact that the stage of the disease is derived from these clinical measurements, including the motricity. The most interesting exception is for the prodromal versus healthy control problem, where our systems performs better. It confirms that brain alteration in PD's prodromal stage is prior to and globally more informative than any motor symptoms, and seem to be correlated with the apparition of non-motor symptoms. Secondarily, it indicates that information from subcortical structures shape alteration is somewhat orthogonal to the clinical motor symptomatology of the patients, and that both should be taken into account when monitoring the evolution of the disease. This conclusion, coherent with the literature [140] [141], is supported by the results presented in Figure 3.4, which shows that motor symptomatology of the patients is largely orthogonal to striatal shape alterations.

Using only subcortical shape analysis, our system shows, with each structure, consistent results for classifying PD patients from healthy controls, whether they are on early or advanced phases, contrary to most of the literature presented in the Introduction, where the results are subtler, and often contradictory [141], [142]. We are also the first to successfully differentiate healthy controls and Prodromal subjects using subcortical shape analysis. The main difference between our study and the current state of the art is that we are using larger cohorts and more advanced machine learning algorithms, which consider non-linear relations along all the surface mesh, whereas the literature usually considers linear relations, one point at a time. This observation shows that subcortical shape alteration is complex and non-linear, and that a supplementary methodological effort has to be done when working with subcortical imaging data in this context.

An unexpected result from our study is that Prodromal patients seem to be easier to distinguish from healthy controls than early PD patients. Although we don't have any clear explanation to provide, this observation illustrates that subcortical shape alteration is not uniform in time.

In term of compression, all four structures displayed the same expected behaviour (Figure 3.3), although right structures appeared to be more easily compressed. This observation is not linked to the volume of the segmentations on the atlas. We cannot be

sure if right structures have a more coherent shape deformation pattern, or if this observation is caused by a unexpectedly noisier output from the segmentation, registration and vertex-wise boundary displacements on the left structures.

We also tried a non-linear data compression method using a neural network and more specifically deep stacked residual autoencoder. In addition, we used feature selection with the k-best algorithm using different correlation scores. However, neither of these approaches yielded consistent better performance while greatly increasing its complexity, in the case of neural networks.

It should be noted that the different stages of PD exist in a continuum, that is, there is no clearly defined, objective delineation between stages. In addition, the Prodromal cohort do not all uniformly progress to full symptomatic PD, although it is increasingly being seen as a common pre-cursor [121]. This has a distinct effect on the accuracy of the methods using either motor symptomology or striatal morphology, specifically for problems that can be seen as 'adjacent' such as classifying between early-stage and late-stage PD, or between healthy controls and the Prodromal cohort.

Our system was able to predict MDS-UPDRS3 score of various PD subjects. The performance is weak, yet significant ($p < 0.001$) as shown in Table 3.8. To us, it indicates that our methodology can be used in theory to predict the motor symptomology of a patient based on striatal morphology, but that they are either not strongly correlated or causally connected. The fact that cohort classification is an easier task than motor symptomatology prediction also supports our conclusion that PD striatal shape alteration is orthogonal to motor symptomatology.

3.5.1 Future Work

By observing that the shape displacements of some structures are more relevant in some problems than others, we hypothesized that striatal structures would deform non-homogeneously through time. We attempted to find a pattern in the spreading of relevant clusters of points through the evolution of the disease, but couldn't find anything reliable. Further study would be needed in this direction. Additionally, we could not find a clear interpretation regarding the most relevant structures of every problem.

In this extent, the longitudinal observation of a cohort of patients would be a good follow up for this study. Indeed, tracking patient's disease on several years could allow to conclude more reliably on patterns in striatal shape alteration by removing the inter-population bias.

A second important direction would be to correlate subcortical shape alterations and the severity of non-motor symptoms, such as cognitive and neuropsychiatric decline. So far, most of the literature, including this manuscript, focuses essentially on motor symptomatology. Finding biomarkers hinting at the severity and the speed of cognitive and neuropsychatric decline would allow for greatly improved medical care, as the choice of treatment would benefit from this information. The ability of our system to classify healthy controls and prodromal subjects, two cohorts for which the motor symptomatology difference is marginal, highlights the relevance of subcortical shape alterations for non-motor symptomatology. Present work could be extended to study other subcortical structures, from other imaging sequences than t1-weighted MRI especially for a better understanding of Prodromal Parkinson's disease with its more heterogeneous symptoms.

Finally, methodologically speaking, our study suffers from some biases. First of all, the sizes of the cohorts are very uneven, which keeps us from directly comparing the relevance of shape displacement for different diseases stages. Secondarily, PPMI being a multi-centre program, the imaging acquisition methodology and devices are not consistent. Although the registration and segmentation processes are designed to be somewhat agnostic to these differences, there is a possibility that they could have impacted the output of the shape analysis pipeline, and thus bias the results. Nonetheless, this study being mostly preliminary and comparative, the observations drawn in this paper are still relevant to the scientific community and give some insight into the progression pattern of PD.

3.6 Conclusions

In this paper we presented a pipeline able to successfully differentiate populations at different stages of PD, solely from the morphological shape alterations of the bilateral caudate nucleus and putamen. This is an important conclusion for the PD research, as MRI is only marginally used as a bio-marker for the disease, compared to clinical biomarkers. Imaging biomarkers also give more direct insight into disease progression, giving targeted information about the patient's neuroanatomy, as opposed to the more global behavioural view attained through measuring manifested motor symptoms. This is crucial for the understanding of prodromal and early stages of PD in which symptoms are not confined solely to motor performance.

Our analysis framework was constructed as a series of binary classifiers distinguishing between different stages of PD. Subjects with advanced PD are reliably distinguished

from the other cohorts, with a high sensitivity and specificity. Distinguishing between more similar stages of PD has lower performance, but still outperforms diagnosis based solely on symptomology indicating that shape alterations occur early in the disease and progress over time. We show that striatal structures shape displacements can be reliably used as a diagnosis bio-marker for PD, even with relatively simple and well-validated machine learning tools.

The performance is more limited in distinguishing between the healthy control, prodromal PD and early PD groups. Yet, the results are significant, and we show that subtle shape displacements exist, are detected and are exploitable with this methodology. Subcortical shape displacement can thus be used as a staging biomarker. We notably showed, for the first time, that striatal structures shape displacements allow to diagnosticate PD's prodromal stage.

We showed that each of these structures are informative in a different way, as the best performance is given when using them all as input. Thus, the shape displacements of each structure carries different relevant information. Then, we investigated on the informativeness of left and right structures, and concluded that the starting side of the disease is directly correlated to it.

Finally, we shown that our system is hardly capable to predict motor symptomatology from striatal shape displacement, showing that both are either poorly, or difficultly correlated.

Acknowledgments

Part of the data used in the preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI) database (`www.ppmi-info.org/data`). For up-to-date information on the study, visit `www.ppmi-info.org`. PPMI - a public-private partnership - is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including AbbVie, Allergan, Avid Radiopharmaceuticals, Biogen, BioLegend, Bristol-Myers Squibb, Celgene, Denali Therapeutics, GE Healthcare, Genentech, GlaxoSmithKline, Lilly, Lundbeck, Merck, Meso Scale Discovery, Pfizer, Piramal, Prevail, Roche, Sanofi-Genzyme, Servier, Takeda, Teva, UCB, Verily, Voyager Therapeutics and Golub Capital.

Chapter conclusion

Brain anatomy is greatly affected by PD, and the state of alteration is a good indicator of the disease progression, but this has been poorly quantified in the literature so far. In this chapter, we proposed a pipeline to extract and compress the entire striatal shape displacement field, resulting in an informative vector that can be used to predict the state of deterioration of the patient's brain. The encouraging results obtained in this chapter show that our method is a great addition in input of our patient-specific CDSS for the prediction of DBS clinical outcomes. We expect that the conjunction of striatal shape alteration (as extracted and compressed by the method proposed in this chapter) and clinical testing (as compressed and curated using method proposed in Chapter 2) could provide enough information about the patient's disease to allow for accurate pre-operative prediction of the success of DBS interventions.

Chapter 4

PassFlow: Patient screening support workflow

This chapter is the third and last on constructing a patient-specific CDSS for the prediction of DBS clinical outcomes from pre-operative information. It takes advantage of the methods presented in Chapters 2 and 3 to transform respectively the results of clinical testing and the T1-MRI scans to readily usable vectors. In this chapter, we will evaluate the ability of an SVM regressor, which takes in input these vectors and demographics, to predict various clinical outcomes of DBS on a retrospective cohort of PD patients, on a pure data-driven fashion. Altogether, this pipeline represents a first version of a core prediction system for a non-knowledge-based CDSS for assisting the screening phase of a DBS, answering a critical problem in the DBS workflow.

This chapter has to be submitted. [1].

1. Peralta, M., Haegelen, C., Jannin, P. and Baxter, J. (2020). Multimodal Pre-Operative Biomarkers Can Predict Deep Brain Stimulation Outcomes. (to be submitted).

4.1 Abstract

Deep Brain Stimulation (DBS) is a proven therapy for Parkinson's Disease (PD), leading most of the time to an enhancement of motor function. Nonetheless, several undesirable side effects can occur after DBS, which sometimes worsen the quality-of-life of the patient. For this reason, the clinical team have to carefully select the patients on whom to perform DBS. Over the past decade, there have been some attempts on relating pre-operative data and DBS clinical outcomes, but most of them solely focused on the motor symptomatology and considered few pre-operative bio-markers. In this paper, we propose a machine learning based method called PassFlow able to predict a large number of DBS clinical outcomes for PD, in order to allow clinicians to better perform patient selection for DBS and to better communicate potential outcomes to patients. PassFlow uses pre-operative clinical, morphological, and demographic modalities, was validated on two databases, showing correlation coefficients as high as 0.71.

4.2 Introduction

Parkinson's Disease (PD) is the second most common neuro-degenerative disease, affecting 1% of the population over 60 years old [7]. The causes are vastly unknown with multiple symptoms, and there is no way to halt the progression of the disease. However, there are therapies which enhance the quality of life of the patient, notably by alleviating motor symptomatology. Nonetheless, even if PD is considered primarily a motor disease, the patient is also impacted by other types of symptoms, such as dementia, depression or apathy [152]. A now common therapy for PD is DBS, which is a neurosurgical procedure consisting in implanting one or two electrodes in order to electrically stimulate deep anatomical structures. The most common targets are the STN, the GPi and the VIM. A reason why DBS is an increasingly common therapy is that, most of the time, it gives better and more stable motor outcomes than pharmaceutical therapies alone. Nonetheless, as every surgery, the procedure involves some potential risks, such as intracranial hemorrhage [153]. Moreover, it is also known that DBS can aggravate other symptoms due to the possible stimulation of adjacent structures and white matter tracts. The literature has, for example, suggested declined verbal fluency [154] [155] or loss of verbal memory [155] can result from the stimulation. This overall impairment of the neuropsychological and cognitive functions can, in the end, lower the quality-of-life of the implanted patient [152]. The great heterogeneity of PD means that the tasks of monitoring and treating appropriately an individual patient are complex [156]. Altogether, these factors make the patient selection crucial, especially in the STN [157]. To date, it is still quite difficult for the clinician to anticipate the post-operative effects of DBS. Even if major guidelines have been defined [158], there is no way to know what could be exactly the positive and negative effects, especially long-term, of a DBS, keeping from consistently make the best therapy decision for every patient [159], and there is an open pressing need of addressing this problem [160].

Clinical outcomes of DBS are affected by the stimulation site, electrical parameters, and by the patient's preoperative clinical state. Only the later will be considered in this paper, as we focus solely on patient selection, and not on electrode placement nor the fine-tuning of stimulation electrical parameters, which are orthogonal intra-operative and post-operative problems.

There have been promising works in predicting the effects of DBS prior to any operation by using pre-operative data modalities. First of all, demographics have been ex-

tensively considered. Jaggi *et al.* [161] showed that age and disease duration are both predictive factors for motor improvement of patients with STN-DBS. More recently, Farrokhi *et al.* [90] showed that the age at surgery is an important predictor of post-operative complications. Additionally, patients with a good levodopa response are often good candidates for STN-DBS [162].

Imaging data, particularly MRI has also been used. It has been shown that the T2 relaxation time of the STN of PD patients is inversely proportional to their UPDRS1 score [163], and also is a predictor of the motor efficiency of DBS [164]. Ballarini *et al.* [165] were able to predict weak medication responders from gray matter density maps. Weinkle *et al.* [166], on another hand, observed that neither white matter lesion volume nor cortical atrophies are great predictors for cognitive changes after a STN-DBS.

A few recent attempts have combined the information arising from several modalities. Habets *et al.* [83] used logistic regression to successfully predict weak motor responders for STN-DBS from pre-operative demographics and clinical tests, such as motor and neuro-psychological assessments. Shamir *et al.* [167] developed a linear function linking several multi-modal variables, such as age at surgery, VTA within the target, and levodopa response to the post-operative UPDRS3 score, achieving satisfying performance.

These attempts demonstrate to us that the clinical outcomes of DBS are highly multifactorial, and that information can lie in multiple data modalities. An efficient post-operative predictive workflow should thus be able to take into account several modalities simultaneously, in a simple and robust manner. The fact that the solution appears to lie in many different, complex and high-dimensional modalities makes this prediction task complex for a human, but suitable for machine learning algorithms [159].

One lacking point from the literature is the diversity of predicted scores. Indeed, as seen in this introduction, a great majority of the efforts have been done towards predicting the motor outcomes, often by the mean of the UPDRS3 score. This is problematic in the sense that, as already stated, the successfulness of a DBS is a trade-off between desired enhancement of the motricity, and occurrence of undesirable side effects. Therefore, a post-op clinical outcome predictive workflow would greatly benefit from being able to predict every relevant clinical scores, such as neuro-psychological ones [152].

Finally, many works have simplified the problem by binarizing it, e.g. by defining a threshold between a 'good outcome' and a 'bad outcome' [165] [83] [58]. To us, this approach lacks of nuance, and limits the impact of such systems. Moreover, the threshold should be left to the clinicians and the patients on a case-by-case basis, and not hard-

coded in the predictive system. The most the predictive system should do is to predict the post-operative scores as accurately as possible, in order to provide clinicians with additional decision making support.

Contributions

In this paper, we propose a multi-modal, machine learning-based workflow, referred as PassFlow, able to predict the clinical outcomes of DBS for PD, by using pre-operative features. PassFlow was able to significantly predict 63 out of 82 different clinical outcomes with correlation coefficients up to 0.71, outperforming a linear baseline. Our workflow consists of a custom supervised ANN, which extracts features from pre-operative clinical scores in the form of a low-dimensional vector. Then, we append additional features such as demographics, and compressed striatal shape displacements to this vector, before making the final predictions of the desired post-operative clinical score with a linear SVM. We validate PassFlow on two databases and compare the performance on the motor scores predictions for the two databases.

4.3 Materials and Methods

4.3.1 Data

For this study, we used a cohort of PD patients who undergone a DBS at the Rennes University Hospital. It consists of three modalities of data: the clinical data, the imaging data, and additional information, such as demographics.

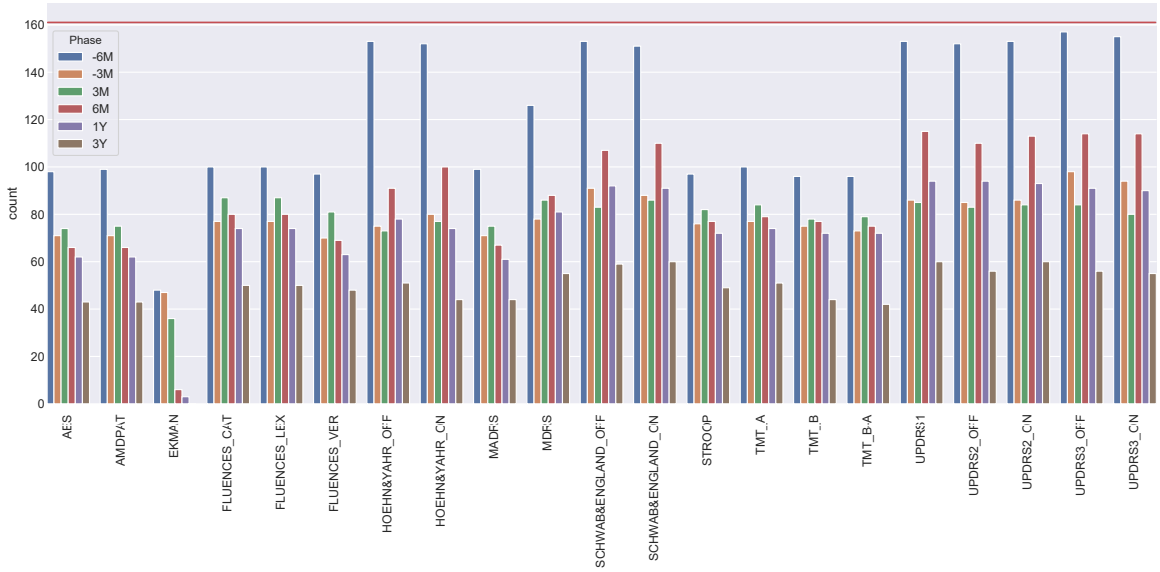


Figure 4.1 – Number of patients for each pre-operative and post-operative clinical scores, for the Rennes DB. Red line shows the number of patient (161) for which there is at least one clinical score.

The Rennes DB contains a total of 196 patients with electrodes implanted in the STN, the GPi or the VIM (characteristics of this cohorts are presented in Table 4.2).

This cohort of patients suffers from missing modalities and missing values within each modality. Indeed, for clinical and additional information modality, data vectors may not be complete. For example, there were two pre-operative visits only for patients undergoing a STN-DBS, and only one for the other targets. Table 4.1 presents the distribution of totally missing, complete and incomplete modalities. Figure 4.1 shows the number of values for each pre-operative and post-operative clinical score.

Clinical data Patients had one or two pre-operative visits (approximately six and three months before the surgery, respectively referred as '-6M' and '-3M'), and up to four post-operative visits (approximately three months, six months, one year and three years after

Modality	Clinical						Imag.	Add.
Phase	-6M	-3M	3M	6M	1Y	3Y		
Complete	0	23	18	1	1	0	180	133
Incomplete	150	83	81	124	103	64	0	28
Missing	46	90	97	71	92	132	16	35

Table 4.1 – Distribution of the modalities of the Rennes cohort.

Age at surgery	Gender (F/M)	Target (STN/GPi/VIM)
59.9 (± 8.1)	68/94	89/54/19

Table 4.2 – Available demographics and surgical target information on the Rennes cohort.

the surgery, respectively noted '3M', '6M', '1Y' and '3Y'). We did not include later visits as the number of data points available were too low.

At each visit, the patients completed the following clinical tests: AES, AMDP-AT scale, Ekman 60-faces test, Categorical Fluency, Lexical Fluency, Verbal Fluency, Hoehn and Yahr (dopa on and off), Montgomery-Asberg Depression Rating Scale (MADRS), Mattis Dementia Rating Scale (MDRS), Schwab and England (dopa on and off), Stroop Test, Trail Making Test A (TMT_A), Trail Making Test (TMT_B), Trail Making Test B-A (TMT_B-A), UPDRS part 1, UPDRS part 2 (dopa on and off) and UPDRS part 3 (dopa on and off, stim off for pre-operative phases and stim on for post-operative phases).

Imaging data Each patient had one preoperative 3T T1 and T2-weighted MRI scan (1mm x 1mm x 1mm, Philips Medical Systems). The T2-weighted MRI scans, were not used for this study.

Additional information As additional information, we collected the patient gender, age at surgery, target (STN, GPi, VIM) and laterality of surgery (left, right or bilateral).

PrediSTIM database Additionally to the Rennes database, we validated PassFlow with data arising from the PrediSTIM protocol, which is a multi-center national protocol centralized by the Lille University Hospital (France), aiming at finding predictive factors of the outcomes of STN-DBS, especially the quality of life. To date, 700 patients have been recruited on this protocol, that started in November 2013. Patients of the PrediSTIM database had one pre-operative visit, and 2 post-operative ones one and three years after surgery. At each visit, patients from the PrediSTIM protocol completed the following

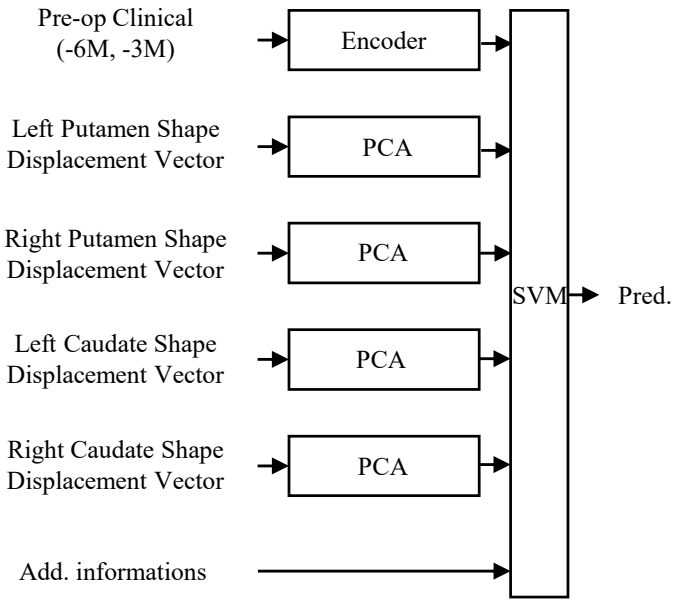


Figure 4.2 – Schema PassFlow. An ANN (PassNet) compresses pre-operative clinical data, and PCAs compress shape displacement vectors of four structures. All these data vectors, with demographics, are fed to an SVM regressor which makes the post-operative score prediction.

clinical tests: Total non motor fluctuations evaluation (V0: dopa off stim off, dopa on stim off, V1: dopa off stim off, dopa off stim on), Miami hallucination scale (V0, V1, V3: total, visual, auditive, somesthesic, gustatory, olfactive), total PDQ39 (V0, V1, V3), CGIS (V0, V1, V3), SCOPA-PS (V0, V1, V3), total MDS-UPDRS part 1 (V0, V1, V3), total MDS-UPDRS part 2 (V0: on/off dopa V1, V3: on stim, on/off dopa), total MDS-UPDRS part 3 (V0: dopa off; 15, 30, 45, 60, 90 and 120 minutes after dopa on, V1: off dopa off stim, on dopa off stim, off dopa on stim, on dopa on stim), total MDS-UPDRS part 4 (V0, V1, V3), Hoehn and Yahr (V0, V1, V3: on, off), Schwab and England (V0, V1, V3: on, off). Additional patient data, for the PrediSTIM database, consists in the gender, the age at surgery, a Boolean indicating if the disease if familial, and a Boolean indicating if the disease is genetic.

4.3.2 Proposed Method

The proposed method, referred to as PassFlow (for Patient Screening Support work-Flow), consists of an SVM, with a linear kernel, getting a compressed clinical data vector, four compressed striatal shape displacement vectors, and additional information such as

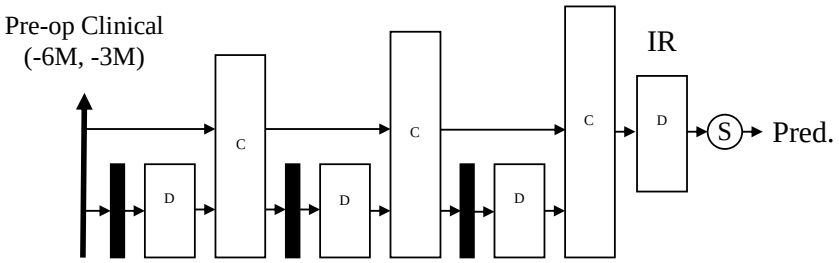


Figure 4.3 – Schema of the ANN structure used for clinical data pre-processing. Blocks 'D' are densely connected layers with ReLU activation, blocks 'C' are concatenation layers, black-filled blocks are dropout layers with a drop rate of 0.1, neuron 'S' is a densely connected neuron with sigmoid activation.

demographics and surgical target. This additional information has not been treated like other clinical data, since they rarely had missing and were primarily categorical, which our current clinical data handling method did not support.

The methods to compress the clinical data and extract the compressed striatal shape displacement vectors are presented in the following sections.

Imaging data processing

Out of the T1-weighted MRI, we extracted four striatal shape displacement vectors, following the methodology presented by Khan *et al.* [145] which uses deformable registration to extract the bilateral caudate and the bilateral putamen in the patient MRI and compare their shape against those in the MNI PD25 Atlas [151]. The shape deformation field sampled at the surface of each of the four structure were then re-organized as a 1D vectors.

As these vectors were high-dimensional (several thousand values), we compressed these vectors separately using a PCA. The number of principal components kept was optimized with the HPO process presented in Section 4.3.5, for each post-operative clinical score.

We chose this methodology to extract bio-markers from T1-weighted MRI because we previously shown their interest as staging bio-marker, and their correlation with the patient's UPDRS3 score [168].

Clinical data processing

Clinical test results from normal clinical routine present two problems for downstream analysis: they are high-dimensional, and they suffer from missing data. We previously pub-

lished an extensive study on compressing clinical data by appropriately handling missing data, and proposed a custom deeply-learned autoencoder that outperformed linear baselines such as PCA [169].

In the current study, we kept the same ANN structure, but instead of training it unsupervisedly, we only took the encoder part and added a sigmoid neuron after the bottleneck to perform regression, and therefore train the encoder in a supervised manner.

This ANN structure is presented in Figure 4.3. The input bias-only layer allows a more advanced missing data imputation strategy than regular zero imputation or mean imputation: the zero-imputed pre-operative clinical data vector is given as input to the ANN, as well as a mask vector indicating where the missing values are. From these two vectors, the bias only layer adds a learnable bias to the input vector at places where values are missing. This way, the optimal imputation value for each clinical score is learned in a supervised manner. Following this data imputation layer, the ANN consists of a series of dense layers followed by a dropout layer, the output of which is concatenated to the input for the successive layers.

The shape of the ANN, e.g. the depth and the number of dense neurons per layer, has been optimized with the HPO process presented in Section 4.3.5. The best shape was defined as the one giving the highest mean regression correlation score for all post-operative clinical scores. Therefore, the same ANN topology was used for every post-operative score.

Other hyper-parameters, such as the learning rate, the batch size and the internal representation size have been optimized for each post-operative clinical score separately.

Once the ANN has been supervisedly trained, we removed the sigmoid neuron and used the output of the internal representation as clinical features. One training has therefore to be done for each post-operative clinical score to predict.

4.3.3 Accuracy and loss metrics

We used coefficient correlation R as a comparison metric between methods, and for HPO. For the regression task, ANN have been optimised on Mean Squared Error (MSE).

4.3.4 Training and validation

For each test, we used a stratified 50-fold CV. Each model has been trained 50 times separately, using one fold as a validation data and the remaining forty-nine as training

data. If a clinical score had less than 50 samples, a Leave-one-out CV (LOOCV) has been used.

4.3.5 Hyper-parameter optimization

The hyper-parameters of each classifier have been optimized for each CV fold using Bayesian optimization with a Gaussian process as a surrogate model and expected improvement as the criterion. The number of points tested was the square of the number of hyper-parameters to optimize plus one.

4.3.6 Software environment

We used Python 3.6, with Keras (version 2.2.4, TensorFlow version 1.12.0 as a backend) for ANN implementations, and Scikit-learn (version 0.21.1) for other machine learning implementations.

Method	Mean R	Max R	Sig.	Not sig.	Discarded
PassFlow	0.32 (±0.20)	0.71	63	19	2
Linear	0.25 (±0.23)	0.68	47	35	

Table 4.3 – Statistics regarding the coefficient correlations on clinical score predictions, with our proposed method PassFlow and the linear baseline.

4.4 Experiments

We conducted an experiment, which consisted in studying the correlation coefficients (R) between the predictions and the ground truths for each post-operative clinical score of the Rennes database. In a second time, we reproduced the work on the PrediSTIM database.

Results were obtained by running 50-fold CV for each post-operative clinical score with at least 10 values, after having done an HPO. This HPO allowed to find the optimal batch size, number of training epochs, number of principal components kept for striatal shape displacement vector, and the size of the intermediate representation for clinical data.

PassFlow results for all scores

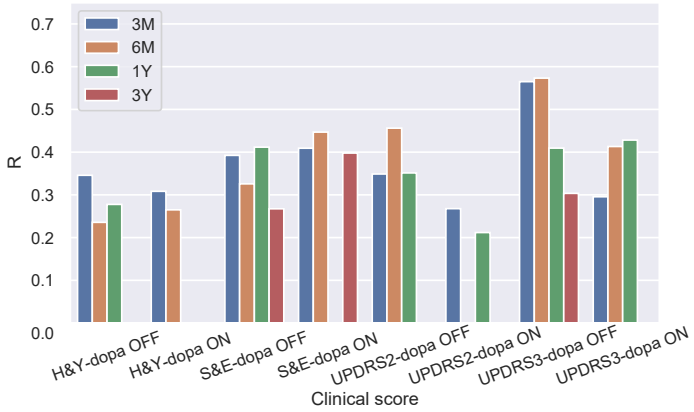
Figure 4.4, shows the correlation coefficient of PassFlow, between predictions and ground truths, for each post-operative clinical scores. Figure 4.4a shows the motor scores, Figure 4.4c shows the cognitive scores and Figure 4.4b shows the behavioral scores.

As an example, Figure 4.5 shows the predictions against the ground truths for the score the most accurately predicted by PassFlow, which is lexical fluency three months after surgery.

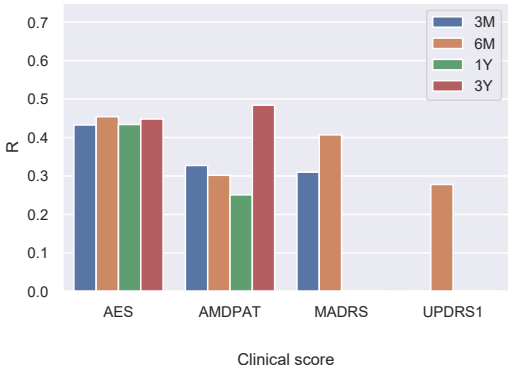
Comparison against the linear baseline

Comparison with all scores

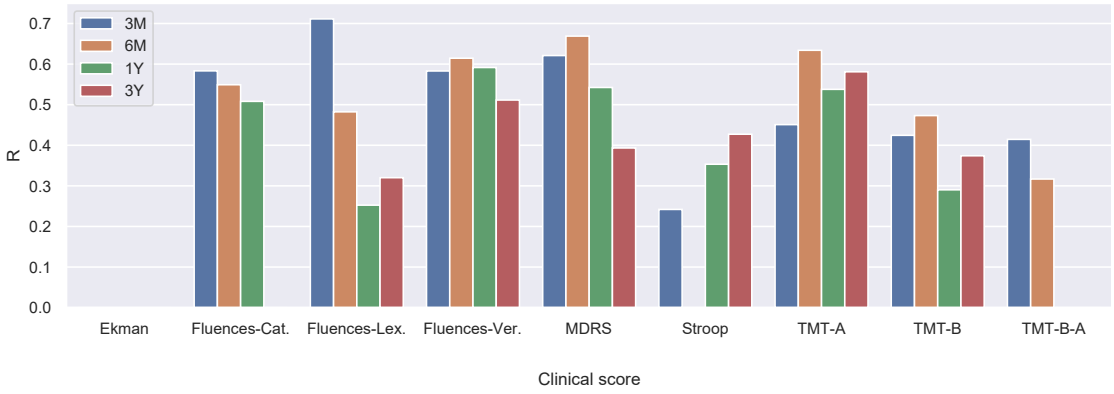
Figure 4.6 compares the distributions of correlation coefficients of the predictions on each post-operative clinical score (provided that there are more than 10 patients), for PassFlow and a linear baseline, which is a simple linear regression taking as input the pre-operative values (at -6M and -3M) of the clinical score being predicted.



(a) Correlation coefficient of PassFlow predictions for motor tests.



(b) Correlation coefficient of PassFlow predictions for behavioral tests.



(c) Correlation coefficient of PassFlow predictions for cognitive tests.

Figure 4.4 – Correlation coefficient (R) of PassFlow predictions for each post-operative clinical scores. For clarity purposes, we split the scores between motor scores, behavioral scores and cognitive scores.

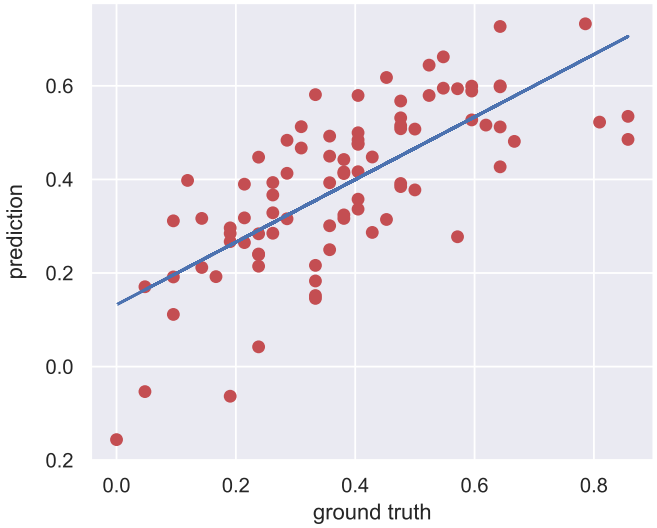


Figure 4.5 – Scatter plots of the prediction of lexical fluencies, 3 months post-surgery, with our proposed system PassFlow.

Table 4.3 synthesizes the results on all scores, showing the number of tests for which we have statistically significant results and statistics on R values obtained, for both PassFlow and the linear baseline. Clinical scores with less than 10 patients were discarded.

Based on these results, we found a significant correlation between the performances of PassFlow and the number of patient for each score ($R = 0.24, p = 0.034$), as shown in Figure 4.7, whereas this correlation was not significant for the linear baseline ($p > 0.05$).

Comparison with common significant scores

Figure 4.8 compares the distributions of correlation coefficients of the predictions on the 42 clinical scores where both methods were significant. It corresponds to filtering every score belonging to the 'not significant' column in Table 4.3.

On this subset of clinical scores, both methods perform fairly similarly. The linear baseline outperformed PassFlow for 22 of them, and the average number of patients for these clinical scores was 73.5. On the remaining 20 clinical scores, PassFlow outperformed the linear baseline, and the average number of patients for these scores was 84.3.

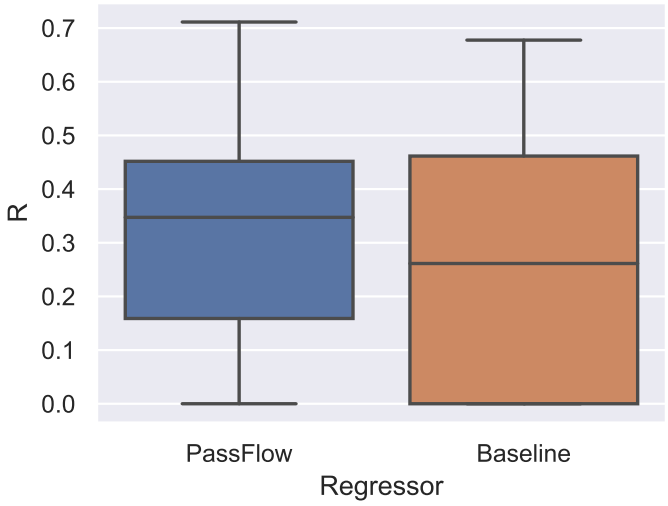


Figure 4.6 – Distribution of the correlation coefficients between the ground-truths and the predictions of both PassFlow and the linear baselines, for all the post-operative clinical scores.

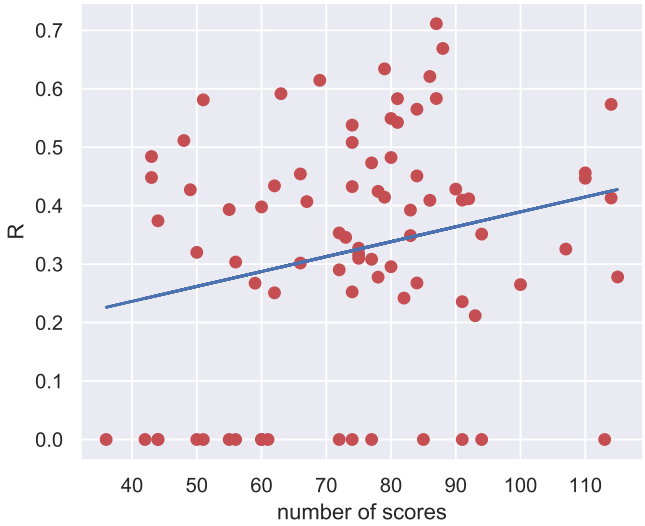


Figure 4.7 – There is a correlation of 0.24 (at $p = 0.034$) between the performances of PassFlow and the number of patient for each score.

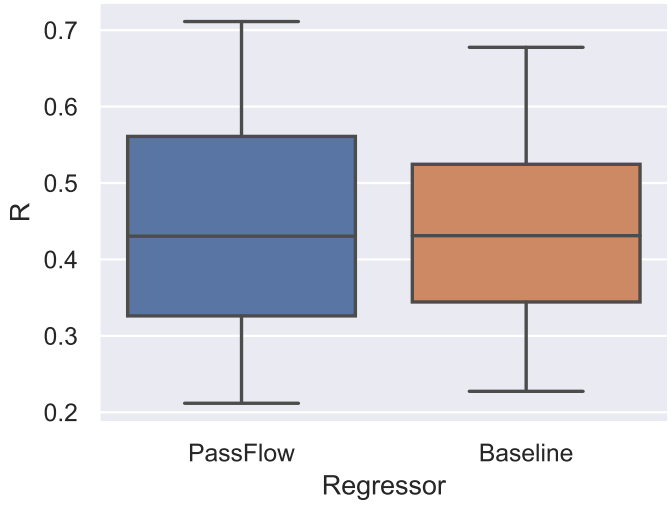


Figure 4.8 – Results distribution by taking common significant scores.

Validation on PrediSTIM database

We used our predictive system to predict an ensemble of post-operative clinical scores (motor and quality of life scores) for the PrediSTIM databases. Results of this experiment are shown in Table 4.4. The results on Rennes database are showed for common motor scores.

Clinical test	Phase	PrediSTIM	Rennes
Common Motor Scores			
MDS-UPDRS2 ON/ON	1Y	**0.3485**	**0.2118**
MDS-UPDRS2 ON/OFF	1Y	**0.2846**	**0.3513**
MDS-UPDRS3 ON/ON	1Y	**0.2951**	**0.4094**
MDS-UPDRS3 ON/OFF	1Y	**0.251**	**0.4282**
Hoehn and Yahr ON/ON	1Y	**0.4833**	**0.2778**
Hoehn and Yahr ON/OFF	1Y	**0.3703**	0.1493
Schwab and England ON/ON	1Y	**0.3100**	0.1477
Schwab and England ON/OFF	1Y	**0.3460**	**0.4117**
MDS-UPDRS2 ON/ON	3Y	0.2115	0.1230
MDS-UPDRS2 ON/OFF	3Y	**0.4414**	0.2546
Hoehn and Yahr ON/ON	3Y	**0.3262**	0.0648
Hoehn and Yahr ON/OFF	3Y	0.2024	*0*
Schwab and England ON/ON	3Y	**0.3047**	**0.2674**
Schwab and England ON/OFF	3Y	0.1931	**0.3979**
Average (± std)		0.31 (± 0.085)	0.26 (± 0.15)
Additional Motor Scores			
MDS-UPDRS3 OFF/ON	1Y	**0.3862**	
MDS-UPDRS3 OFF/OFF	1Y	**0.3803**	
MDS-UPDRS4	1Y	**0.2493**	
MDS-UPDRS4	3Y	**0.3533**	
Quality of life score			
PDQ39	1Y	**0.5024**	
PDQ39	3Y	**0.3812**	

Table 4.4 – Correlation coefficient between the predicted and true post-operative clinical scores with those significantly greater than 0 shown in **bold** (with a statistical threshold of $\alpha = 5\%$). The predicted clinical scores are divided into three subgroups: the motor scores which are available on both databases, the clinical scores only present on the PrediSTIM dataset, and quality of life scores. ON/OFF stands for stimulation ON, dopa OFF.

4.5 Discussion

Our workflow, PassFlow, was able to significantly outperform a linear baseline, as shown in Figure 4.6. Table 4.3 shows that PassFlow presented a higher mean and maximum correlation coefficient across all the post-operative clinical scores to predict, and was able to predict more clinical scores than the linear baseline. This result indicates that PassFlow successfully took advantage of the extensive pool of features used, i.e. the compressed clinical pre-operative state of the patient, imaging bio-markers and additional information such as demographics and target structure, allowing it to predict more scores than a linear method considering only one clinical score in input. Nonetheless, those better performances mostly come from the fact that PassFlow was able to make significant predictions for more scores than the baseline. For scores that were significantly predicted by both methods, the performances were similar, as shown in Figure 4.8. We can see that the standard deviation of PassFlow's performances was greater than the linear baseline's ones, denoting that PassFlow has the potential to consistently outperform the baseline, but that training issues remain for some scores.

In this extent, we showed that the number of patients is crucial in term of performances for PassFlow. On the first experiment, we noted a significant correlation ($R = 0.24$ at $p = 0.034$) between PassFlow's prediction performances and the number of patients, a correlation that was not significant for the linear baseline. Additionally, in the second experiment, we showed that PassFlow outperformed the baseline for clinical scores with a larger number of patients, and got outperformed for clinical scores with less patients. Both of these results showed that PassFlow can continue to learn and as room for improvement given a broader or more complete database, where it is not the case for the linear baseline. It is easily explained by the curse of dimensionality, as PassFlow took more features as input, increasing the risk of overfitting for small databases.

To go further, clinical scores at later visits seem to be harder to predict. This can be explained either by the fact that there is a greater variability of the disease as it progresses, or by the fact that there were fewer patients for which scores were available, rendering the training of PassFlow more difficult. Unfortunately, the high variability in the number of datasets available for each clinical score prevents us from drawing further conclusions regarding their relative complexity (that is, questionnaires undergone by a larger number of patients tended to be easier for the algorithm to predict, regardless of the questionnaire's complexity).

We validated PassFlow on a second database, PrediSTIM, and obtained better results on the motor scores than on the Rennes database, as seen in Table 4.4. These better results can be explained by the highest number of patients, and also by a greater homogeneity of the database, as PrediSTIM is only composed of patient stimulated in the STN. On top of motor scores, PassFlow gave good results on the prediction of quality-of-life scores ($R = 0.50$ one year post-surgery, and $R = 0.38$ three years post-surgery). Predicting the patient quality-of-life is one of the most important, yet difficult, aspects of patient selection for DBS electrode implantation. These questionnaires are often about the patient's subjective experience, which makes them significantly more difficult to use in population-wise machine learning contexts. It is therefore worth noting the high correlation coefficient seen for the PDQ39 quality-of-life score. Thus, by predicting not only the more objective motor and cognitive scores, but accurately predicting quality-of-life scores using a relatively simple and robust machine learning pipeline represents a potentially large advancement in how patients are selected for DBS and informed about potential side-effects.

All together, the high number of scores successfully predicted by PassFlow indicates that some aspects of post-operative clinical effects of DBS can be anticipated pre-operatively. For certain scores, like fluencies, most of the variance seems to be explainable by pre-operative features. This may indicate that DBS surgeries are consistent in terms of electrode placements affecting these scores. Nonetheless, most of the predictions are in the low to moderate correlation range. This denotes that methodological improvements, as well as supplementary data collection efforts would be necessary to integrate PassFlow in clinic. That being said, it seems hard to determine a threshold above which the performance of PassFlow could be considered as satisfying. Indeed, PassFlow is trained with a clinical ground-truth which is, by essence, inaccurate. First, this ground-truth is subject to the assessors' subjectivity, or to the patient's own perception of its symptomatology. Second, the clinical state of a patient is not stable. The patient can be on a *good* or a *bad* day when performing the test. Moreover, instead of aiming for unreachable perfect performances, maybe the first objective for a predictive system such as PassFlow would be to surpass human performance. When it comes to predict the clinical outcomes of DBS, human expert performance has, to date, not been quantified.

4.5.1 Future Work

Several aspects of this study could be extended. Firstly, relative importance of each modality could be extensively studied, globally and for each score. Our recent preliminary investigations show that imaging bio-markers could not have a significant impact on the performance, and could even have decreased the performance for some clinical scores. Quantifying the effectiveness of each biomarker would be an interesting follow-up work. This way, the most informative data modalities could be identified, and several ways to extract features from them could be benchmarked, notably for imaging data. To go further, clinical hypothesis could arise from such a data-driven study. Secondly, the current method has been validated on PD, but it could be extended to any other disease which is treated using DBS, provided a large enough database is available. It could also allow for comparing between predictive models, in order to see if there are common risk-factors between the diseases, identifying common DBS effects. Finally, the pool of pre-operative bio-markers could easily be extended, especially regarding imaging data. Indeed, information lay in more imaging data modalities than t1-MRI scans and in more structures than the striatum.

Adding electrical parameters to the model

By adding electrical parameters in input of the model, PassFlow could be extended to an orthogonal clinical problem which is the target choice and electrical parameters tuning. We could expect an enhancement of the results, as DBS post-operative clinical outcomes are function of the pre-operative clinical state of the patient and of the stimulation itself.

A lot of attempts have been done in the past decade to understand the mechanisms of the stimulation, which remain vastly unknown [170]. A first majorly investigated strategy is to consider the stimulation localization and electrical parameters to predict the effects of the stimulation. The results seem more promising when the contact localization and electrical parameters are put in relation with anatomical or functional information. Lalys *et al.* [2] proposed a clustering method linking three motor and five neuropsychological scores with the spatial coordinates of the active contacts, leading to the construction of anatomo-clinical atlases for STN-DBS. Following this direction, Haegelen *et al.* [3] proposed new anatomo-clinical atlases for the STN and the GPi, and used a VTA model to predict the effect of stimulation on four clinical scores at several locations around the target. Nestor *et al.* [171] tried to map the localization of the electrodes with the motor

outcomes of the surgery, but did not find a significant correlation. Verhagen *et al.* [172] successfully correlated the localization of stimulation contacts with the motor outcomes, by defining the position of the STN with MER. Aviles *et al.* [173] used MRI scans to locate stimulation contacts, and notably showed that a stimulation in the limbic areas of the STN leads to worst motor outcomes than stimulation in the sensorimotor areas. Horn *et al.* [174] used diffusion MRI and a model of the VTA, to build a connectivity profile of the patient, between the stimulation site and other brain locations, and showed that this connectivity profile is a predictor for the clinical motor effects of DBS.

These works present interesting methods to add stimulation related parameters in input of PassFlow. Doing so could allow to quantify in which extent the post-operative clinical outcomes are explainable by known factors, and on which extent there remain to be discovered.

4.6 Conclusion

In this paper, we presented a novel, machine learning based pipeline, referred as PassFlow, to predict a variety of post-operative clinical outcomes of DBS for PD patients. PassFlow took into account various bio-markers, arising from different data modalities, such as compressed pre-operative clinical features and compressed striatal displacement. PassFlow has been able to significantly predict most post-operative clinical scores, showing high correlation coefficients for some scores from pre-operative data only, and that better performances could be achieved using larger databases. It indicates that a lot of clinical outcomes of DBS could be predicted pre-operatively, as PassFlow has been validated without stimulation-related information. By directly predicting quality-of-life scores, this tool can allow clinicians to better perform patient selection for DBS and to better communicate potential outcomes to patients. Finally, PassFlow is flexible, and can be extended to other data modalities, such as other imaging sequences or pipelines. Taken together, PassFlow represents a promising step towards computer-assisted patient screening, reducing the amount of uncertainty clinical teams have in deciding which patients could benefit from DBS.

Chapter conclusion

This chapter presented our patient-specific CDSS for the prediction of DBS clinical outcomes from pre-operative information. Thanks to two pipelines developed in the previous chapters, we transformed our multimodal input data to a readily usable form. Our predictive system, PassFlow, successfully predicted most of the post-operative clinical scores on two databases with satisfying performances, outperforming a linear baseline. PassFlow is a promising and innovative preliminary work on developing a data-driven CDSS to help the clinical team selecting patients to undergo DBS.

Chapter 5

SepaConvNet: Separable-convolution-based convolutional neural network

After having focused on a problem encountered during the inclusion phase of the clinical workflow, we decided to propose a second CDSS to address a challenge encountered during the surgery itself: how to locate, as fast and precisely as possible, the surgical target relatively to the electrode trajectory thanks to microelectrode recordings (MER). For this contribution, we decided to use an approach orthogonal to the state of the art by not relying on feature engineering, but rather using Convolutional Neural Networks (CNN) to analyze the raw signal directly in the form of spectrograms. We hypothesize that such an approach, on top of performing better, could allow for a quicker detection of the structure of interest (here, the subthalamic nucleus (STN)), and could be transferred to other recording conditions easily. On this chapter, we will propose a custom CNN topology and evaluate its ability to locate the STN from one second MER.

This chapter has been accepted at the 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society in conjunction with the 43rd, Annual Conference of the Canadian Medical and Biological Engineering Society, and will be included in the proceedings [1].

1. Peralta, M., Quoc, A., Ackaouy, A., Martin, T., Gilmore, G., Haegelen, C., Sauleau, P., Baxter, J.S.H. and Jannin, P. (2020, July). SepaConvNet for Localizing the Subthalamic Nucleus using One Second Micro-Electrode Recordings. In 42nd Annual International Conferences of the IEEE Engineering in Medicine and Biology Society in conjunction with the 43rd, Annual Conference of the Canadian Medical and Biological Engineering Society.

5.1 Abstract

Micro-electrode recording (MER) is a powerful way of localizing target structures during neurosurgical procedures such as the implantation of deep brain stimulation electrodes, which is a common treatment for Parkinson's disease and other neurological disorders. While MER provides adjunctive information to guidance assisted by pre-operative imaging, it is not unanimously used in the operating room. The lack of standard use of MER may be in part due to its long duration, which can lead to complications during the operation, or due to high degree of expertise required for their interpretation. Over the past decade, various approaches addressing automating MER analysis for target localization have been proposed, which have mainly focused on feature engineering. While the accuracies obtained are acceptable in certain configurations, one issue with handcrafted MER features is that they do not necessarily capture subtler differences in MER that could be detected auditorily by an expert neurophysiologist. In this paper, we propose and validate a deep learning-based pipeline for subthalamic nucleus (STN) localization with micro-electrode recordings motivated by the human auditory system. Our proposed Convolutional Neural Network (CNN), referred as SepaConvNet, shows improved accuracy over two comparative networks for locating the STN from one second MER samples.

5.2 Introduction

Besides the adequate choice of structure, the most determinant factor of the surgery's success is the accurate positioning of the stimulating electrodes relative to the structures of interest. This positioning is currently performed in multiple phases. As pre-operative MR imaging is a routine part of clinical care for this procedure, a surgical planning phase is performed involving the identification of the target and the appropriate craniotomy site on the patient's skull, and the corresponding electrode trajectory. The second phase involves mounting a stereotactic frame onto the patient's skull, re-imaging them and using image registration to find the target co-ordinates in the space of the frame, allowing for the electrode to be implanted guided by the stereotactic frame in the operating room. However, preoperative imaging does not guarantee correct positioning of the electrode during the operation for various reasons including the limited resolution of the images and the potential registration error between the patient space and the stereotactic frame space. Additional deformations during surgery due to brain shift resulting from the craniotomy

imply that the selected stimulation target coordinates determined in the surgical planning phase may no longer correspond exactly with the selected stimulation target location [175]. Lozano *et al.* [176] estimated the frequency of sub-optimal electrode implementation to be 20% for DBS performed solely with the preoperative imaging. Thus, an additional 'fine-tuning' phase is required to ensure a correct implantation.

This 'fine-tuning' stage is performed via 'trial-and-error' when a small amount of stimulation is applied to the electrodes and the effects of said stimulation are used to infer whether or not the electrodes are in the correct position. This process is necessary as the ideal target for stimulation is defined functionally rather than purely anatomically, that is, the ideal target is one in which the stimulation has the desired effect. This process however can be made complicated by the use of different anesthesia regimens during the procedure. Local anesthesia maintains the wakefulness of the patient and thus allows the clinical team to test the clinical effects of various electrode positions more easily, incorporating additional behavioral and guided motor activities. However, performing the procedure with the patient awake can be a source of discomfort for the patient and additional patient movement can be a source of risk. Global or general anesthesia implies that the patient is not awake during the surgery, making the surgery more comfortable and controlled. However, this limits the amount of clinically useful information accessible to the surgical team during the procedure. Regardless of the type of anesthesia used, a second source of information that informs the position of the electrode relative to the targeted structure is considered mandatory either to limit the number of 'trial-and-error' in the case of local anesthesia or to ensure a higher success rate in the case of global anesthesia.

One possibility is to use intraoperative MRI to refine the anatomical localization of the electrode during the surgery, but this can be prohibitively expensive and requires a specially designed operating theatre. An alternative approach is MER, which uses the electrical components of the stimulation electrode to record electrophysiological signals along the trajectory selected prior to the surgery. Using a segment of these signals, the clinical team listens to and/or visualizes the raw MER signal to determine if the micro-electrode is within the surgical target nuclei at the time of recording. Performing this assessment at several depths along the implant trajectory provides the surgical team enough evidence to delineate the borders of the surgical target nuclei. Whether to use intraoperative imaging or MER remains under debate. Lee *et al.* [177] and Liu *et al.*

[178] compared the usage of intraoperative MRI versus MER, showing equivalent clinical outcomes.

We identified two limits of MER that can be leveraged by our approach. Firstly, the MER procedure is time consuming as many depths along the electrode trajectory are to be tested (each depth is an average of 10 seconds in duration). Secondly, this procedure is subjective as whether or not the signal originated from within the desired anatomy is determined by the neuro-physiologist based on their qualitative visual and auditory assessment of the signal. Quantitative criteria used to identify the surgical target are specific neuronal firing patterns and an increase of background noise. However, these criteria are not always obvious and depend on the interpretation by the clinical team, which requires significant training and expertise in this domain.

Contributions

In this paper, we propose a deep learning framework named SepaConvNet (Separable Convolution-based neural Network) for the detection of MER signals arising from the STN, differentiating them from nearby structures, using 1 second MER recordings. This state-of-the-art method shows high accuracy using class balanced accuracy metrics and outperforms ResNeXt, a state-of-the-art CNN architecture. This framework can work in real-time augmenting the clinical teams' capability in terms of DBS electrode implantation for PD.

5.3 Theory and Previous Work

Automating STN localization with MER is an active field of research. Originally, these strategies were based on developing descriptive features characterizing the MER data and applying a classifier to these features. (A detailed review of such methods has been performed by Wan *et al.* [179].) These characteristics fall into two large classes: spike independent features [62], [180], [181] (e.g. power spectral density) many of which can be computed or approximated with an appropriate filter-bank, and spike dependent features [43], [52], [182]–[186] (e.g. standard deviation of interspike intervals). While adding spike dependent features can provide additional information about the signal, it also adds several constraints about signal acquisition [179].

The large amount of work that has been performed on MER feature engineering for the past decade is indicative that optimal features are likely tedious to develop. However,

the fact that trained neurophysiologists can audibly distinguish different MER signals (as it can be in the current clinical workflow) hints that learned features arising from the raw signal itself may provide better outcomes. However, the extent and variety of features investigated demonstrates the hidden complexity of MER signal analysis. To this extent, using a CNN to automatically learn non-linear combinations of convolution filters to address this problem may be considerable easier than using hand-designed ones given the large span of linear and non-linear features that have been previously been considered.

Spectrogram representation of MER

In current clinical practice, MER signals can be interpreted by a trained neurophysiologist by listening to them, indicating that the patterns of interest within these signals are detectable using the human ear. It is well-known in many fields such as perceptual psychology that the ear performs a frequency decomposition of incoming auditory signals and that this decomposition is the data provided to the central nervous system. Spectrograms have previously been used for this application; Karthick *et al.* [187] successfully classified STN MER signals by designing features in a time-frequency domain. Thus, we chose a spectrogram representation of these signals as an initial stream of input features. This representation allows us to provide a more complex and predictive input data representation without the complex feature engineering and can be applied in real-time facilitating its integration into clinical workflow.

Convolutional Neural Networks

Originally developed by LeCun *et al.* [188], CNNs are a type of deep neural network based on convolution layers. These layers consist of several tunable kernels operating on the input space by a linear convolution operation usually followed by a point-wise non-linear activation. Stacking these convolution layers allows the neural network to learn to construct higher level features from the original input space which are more readily used for classification tasks. Following the convolutions layers, typically one or two fully connected layers are added, which use the high level features to perform the classification. During the training process, these kernels and the weights in the fully connected layers are iteratively updated through a process of optimizing a loss function which acts as a surrogate for optimizing accuracy.

ResNeXt

ResNeXt [189] is an extension of ResNet [190], which are both well-known architectures in the computer vision community. Aside from its good performance, the structure of ResNeXt is highly scalable. Indeed, one of the motivations of the original study [189] was to be able to control the shape of the structure of the network with only three parameters, where former well performing CNN models needed almost each layer to be tuned manually. Therefore, it is common to reuse pretrained models and finetune them, but there is no major publicly available CNN model pretrained on MER spectrograms. The ability of ResNeXt to be scalable motivated us to use it as a CNN baseline, to compare the performances of our proposed model.

5.4 Material and Methods

5.4.1 Data acquisition

The MER signals in this study were recorded from 57 PD patients having undergone DBS surgery at London Health Sciences Centre at Western University hospital. The Bens gun was used to advance 3-5 microelectrodes as a unit. Signals were recorded from 10.0 mm above the pre-operatively determined target zero point to 4.0/5.0 mm below the zero-point, the end-point being detectable activity from the Substantia Nigra. The drive was advanced in 1.0 mm increments and 0.5 mm increments within the target nucleus. At each depth, a 10-second recording was obtained once a clean recording was observed (i.e. free from drive noise). The signals were sampled (24kHz, 8 bit), amplified (gain: 10,000) and digitally filtered (bandpass: 500-5000 Hz, notch: 60Hz) using the Leadpoint recording station (Leadpoint 5, Medtronic). The ground truths were determined by the electrophysiologist and neurosurgeon at the time of the recording in the operating room. The database was approved by the local ethics board at Western University (REB #1090485).

5.4.2 Database construction

By observing the signals individually, we noticed that the signals were highly temporally redundant. Usually, the neurophysiologist can be certain of the location of the electrode by listening to several seconds of signal. We hypothesised that a deep learning algorithm can exhibit good performance with far less signal (i.e. only one second) which

could improve clinical workflow by reducing procedure time. Reducing the duration of the signal has three positive effects for the training. Firstly, a reduced input dimensionality makes the deep neural network less prone to overfitting, thus allowing better generalisation performance with a limited number of training samples. Secondly, a smaller input limits the memory amount required to store the input, the gradient, the inference tensors, etc. This smaller memory consumption allows to train the neural network with a larger batch size, and thus train the network more quickly. Lastly, not using the whole 10 second signal allows us to take several shorter windows of each signal, thus increasing the database size.

We took advantage of the third property to balance our databases. Indeed, in our database, there are approximately three times more signals labeled as outside of the STN ('0') than signals labeled as inside of the STN ('1'). To overcome this class imbalance, for each signal labeled as '0' one 1 second window was extracted, were three different 1 second windows were used for each signal labeled as '1', thus balancing the dataset.

To split the data into training, validation/test sets, we used a stratified 5-fold CV. This data splitting technique consists in splitting the data in 5 different folds, with approximately the same number of signals in each fold, with the constraint that all signals coming from a single patient are assigned to the same fold. This constraint is mandatory to ensure that the estimation of network performance is representative of new patients as no patient appeared in both the testing and training sets for either class.

Once the stratified fold splitting is done, 4 folds were used as training data, and the last one as testing/validation data. This process was repeated 5 times in total in order to have every fold used as testing/validation once.

We obtained a database with 9055 one second signals labelled as '0' (outside the STN) and 8255 labelled as '1' (inside the STN), each signal having a length of 24000 samples.

5.4.3 Signal preprocessing

The first step of signal preprocessing was to threshold the signal, in order to limit the effects of non-informative acquisition artifacts. That being done, the temporal input signal is converted into a spectrogram representation. To compute the spectrogram, a Short Term Fourier Transform (STFT) was used, which applies a sliding window to the signal and computes the Fourier transformation for each window. From this Fourier transform, only the magnitudes of the vector were retained. Using STFT leads to a fundamental tradeoff: if the sliding window is wide, the spectrogram will have a higher accuracy in

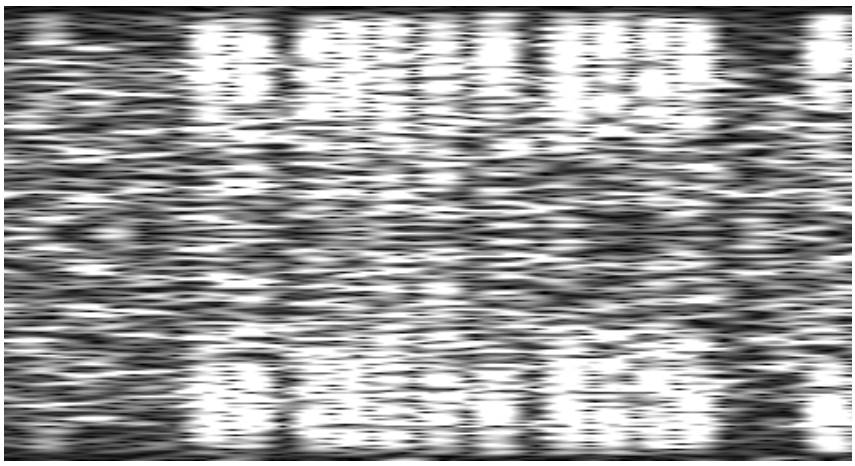


Figure 5.1 – Raw spectrogram of 0.2 seconds of MER signal inside the STN after normalization.

the frequency dimension, but will lose resolution in the time dimension. The STFT was performed with a Hann window of 512 samples width and a hop length of 10 samples, giving us input spectrograms with 257 frequency bands and 2400 time points. An example MER spectrogram is shown in Fig. 5.1.

5.4.4 Proposed Convolutional Neural Network

In the literature, convolution kernels are often 2D as most CNNs are used to address computer vision problems, in which the input space is composed of two spatial dimensions. Spectrogram classification with 2D CNNs has already been successfully performed on music [191], [192] and Doppler radar [193] which present at least some informative frequency shift-invariant properties. However, for MER signals, the two axes of our spectrograms are fundamentally different and there is no practical reason that a pattern detected in the low frequency bands would be equivalent to one in the high frequency bands. Thus we claim that in our use-case, 1D kernels are theoretically more adequate and our input signals are not treated as 2D images, but as 1D vector-valued signal with each element being the intensity in a frequency band.

Furthermore, in MER analysis, the amount of data available is limited. Thus, we replaced traditional convolutions with separable ones in our proposed CNN structure. Separable convolutions have fewer trainable weights and are therefore faster to train for a similar theoretical computational power, allowing to achieve similar results with a lighter network [194]. For small datasets with high dimensionality, a lighter model can limit the

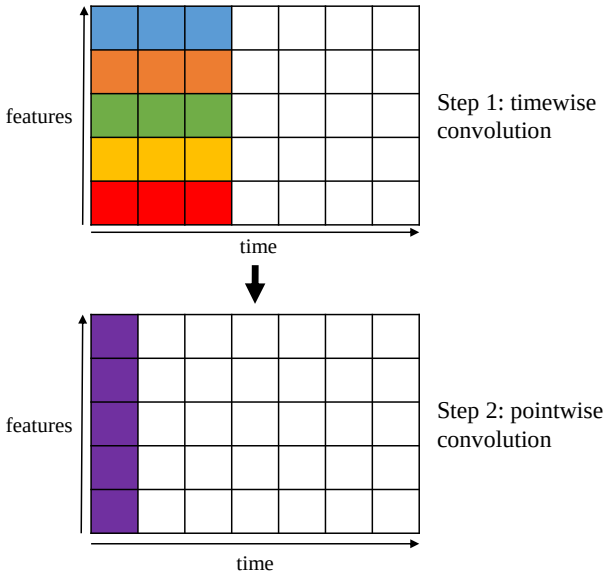
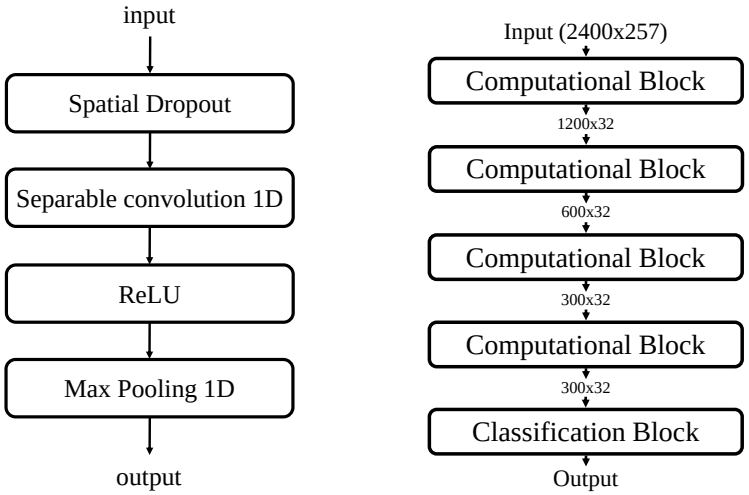


Figure 5.2 – Example of a separable convolution. On the first step, 5 kernels (one for each feature band) of size 3 are convolving along the time axis. The resulting matrix is passed as input of the second step, which performs a convolution along the time axis with N kernel of size 1, mixing all the features together at each timestep. The dimensions of the output matrix are the number of timesteps times N.

impact of overfitting, thus leading to better classification results and training stability. Separable convolution works in two steps. First, a timewise convolution is performed, using K kernels per feature. Second, a pointwise convolution is used in order to mix channels to produce N new features. Given an appropriate choice of K and sufficiently large kernels, this can largely reduce the number of parameters in the total convolution operator. Figure 5.2 shows these two stages of a separable convolution operation with $K = 1$ and $N = 1$ and would output a 1D-vector. In our paper, the hyperparameters K and N have been optimized through a Bayesian process, described in Section 5.4.6.

Our proposed CNN structure, called SepaConvNet (Separable Convolution-based neural Network) is presented in Figure 5.3. It consists of D successive computational blocks followed by a global average pooling layer (along the time-axis) and a fully-connected neuron with sigmoid activation, in order to perform the binary classification. Each computational block is composed of a spatial dropout with a drop rate of 0.1, followed by a 1D separable convolution with kernel size of 15 with ReLU activation, a concatenation layer and a 1D max pooling layer of size 2 and stride 2 on the time axis. All computational



(a) Computational block (b) Complete network

Figure 5.3 – SepaConvNet structure. Figure (a) presents the composition of a computational block. Figure (b) presents the global network, the classification block being an average time-pooling layer followed by a fully-connected neuron with sigmoid activation.

blocks are identical in parameterization, except the last one, which doesn't have a max pooling layer.

5.4.5 Accuracy and loss metrics

As classes remain slightly unbalanced, we used balanced accuracy (Eq. 5.1) to measure the classification performance and to compare methods. Each classifier was trained with class weighting, giving a weight to training sample inversely proportional to the representation of the belonging class in the training set, thus assigning the same total weight to both classes regardless of the number of samples in each.

$$BACC = \frac{1}{2}\left(\frac{TP}{P} + \frac{TN}{N}\right) \tag{5.1}$$

5.4.6 HPO

An HPO has been done using Bayesian optimization, with Gaussian process as a surrogate model, and expected improvement as a criterion, with a 5-fold cross validation. The number of points tested was the square of the number of hyper-parameters to optimize plus one.

Network	BACC	Sens.	Spec.	f1
SepaConvNet	80.9%	75.6%	86.1%	79.2%
ResNeXt 2D	77.3%	70.1%	84.5%	74.9%
ResNeXt 1D	76.8%	66.0%	87.6%	73.4%

Table 5.1 – Mean classification performances for the three methods investigated. 'Inside the STN' is considered the positive class.

5.4.7 Software environment

All neural networks have been developed in Python with Keras (version 2.2.4), with Tensorflow (version 1.12) as a backend. STFT has been performed with Librosa (version 0.7.0).

5.5 Experiment

Our network was compared against two baselines: ResNeXt 2D, which is the standard version of ResNeXt, and ResNeXt 1D, with every 2D operation replaced by its 1D equivalent. Each CNN has been trained and tested with the same 5-fold CV partitions. We used a learning rate of 0.001 for both ResNeXt versions, 0.005 for SepaConvNet, with Adam optimizer and a batch size of 32.

For SepaConvNet, we performed HPO on the parameters D, K and N as presented in Section 5.4.4. For both versions of ResNeXt, we optimized the number of filters in each convolution layer and the depth of the network, and fixed the cardinality to 16. It is to note that the time resolution of input spectrograms had to be decreased for ResNeXt 2D due to an excessive memory consumption.

Results of the 5-fold CV are reported in Figure 5.4 and Table 5.1. Classification results between SepaConvNet and both ResNeXt versions were statistically significant with a paired sample t-test ($p < 0.05$).

5.6 Discussion

The results presented in Section 5.5 confirm our hypothesis that for this particular problem, the spectrograms should be treated as 1D vector signals. Despite the increased capability for parameter re-use, the ResNeXt 2D version did not significantly outperform its 1D counterpart, suggesting that MER signals do not have any significant non-trivial

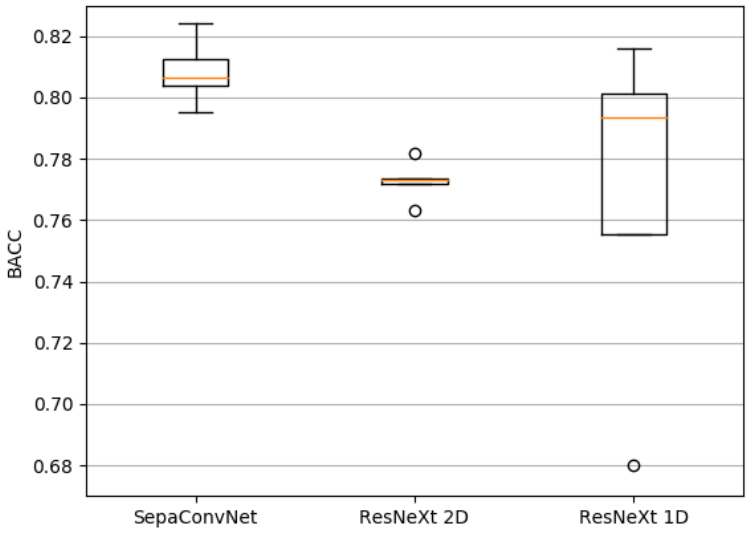


Figure 5.4 – Quantitative results showing the distribution of 5-fold CV balanced accuracies for the three methods investigated.

frequency invariant features (i.e. those in which the shift invariance in the frequency domain would provide a benefit in detecting).

Secondarily, SepaConvNet significantly outperformed both baselines. After HPO, our proposed CNN used fewer trainable parameters than both 2D and 1D ResNeXt (resp. 16.752, 373.953 and 240.449). This decrease in parameters however is not a sign that the ResNeXt networks were overfit as HPO did cover configurations with lower parameterization. However, it does indicate that generic but lighter weight models (largely using separable convolutions) can detect important features in neurological MER signals.

Future work

For future work, we would like to test SepaConvNet on other MER databases to determine if these networks can transfer well to different institutions with different MER hardware.

We would like to investigate on the impact of input signal length on the performance of the system. We would also like our system to allow arbitrary signal lengths to be used, in order to enhance the system ability to work in real time.

Finally, we would further define the classification block of SepaConvNet. Currently, we are using time average pooling followed by a fully-connected neuron. We hypothesize that replacing it with a recurrent layer could enhance the classification performances, as the network could learn the optimal time-pooling strategy.

5.7 Conclusions

We proposed a custom CNN architecture, referred as SepaConvNet, based on 1D separable convolution to analyze STFT spectrograms of MER signals. We experimentally showed that, for MER spectrogram analysis, the SepaConvNet structure outperformed one state-of-the-art structure for image analysis. Our results suggest that deep learnt spectrogram-based features allow the achievement of comparable, or more optimal classification results, than commonly used time domain features. Our system can accurately detect the STN with one second of MER signal. An automatic analysis of MER signal with our system could thus greatly enhance the DBS procedure by significantly reducing its duration.

Chapter conclusion

With this chapter, we proposed a solution to address a pressing clinical problem encountered during the surgery, which is how to locate the targeted structure relatively to the electrode trajectory, and we demonstrated accurate predictions by using only one second of signal. On top of its clinical interest, this chapter also demonstrates the interests and power of data-driven methods, as it exhibited results comparable to those of state of the art feature engineering-based methods which required more than a decade of work and expert knowledge.

Chapter 6

Discussion

In this thesis, we proposed two CDSSs to address two challenging, crucial and concrete clinical problems encountered during two phases of a DBS workflow (the pre-operative phase, with the patient selection problem, and the surgery phase, with the target location problem), as seen in Figure 1.2. In this chapter, I will summarize the contributions of this thesis, focusing on their discussion sections, limitations and perspectives. Following it, I will conclude on the hypotheses stated in the introduction regarding the relevance of data-driven methods and ANNs in DBS. Finally, upcoming challenges and perspectives of data-driven methods in DBS will be presented in a third section before concluding.

6.1 Chapter-by-chapter discussion

In this section, I will review the contributions presented in the previous four chapters: I will summarize their objective, their results, their limitations and the future work considered.

Chapter 2: PatiNAE

In Chapter 2, we proposed a custom deeply-learned AE, referred to as FCAE, able to compress information arising from patient clinical tests and questionnaires while being robust to missing values. The objective was to propose an efficient way to synthesize and curate this rich source of information regarding the progress, the form and the particularities of the patient's disease. Indeed, clinical tests are not readily usable for downstream analysis because they are numerous, multimodal, noisy and suffer from missing values. One way of addressing these issues is to project the high-dimensional input data into a smaller latent space, which is robust to missing values. Our data-driven method takes advantage of 1) the ability of AEs to find complex, non-linear patterns in the input data and 2) a loss function tailored to train the model explicitly to impute missing values. We

showed that FCAE consistently outperformed two linear baselines on data imputation and reconstruction for different latent space dimensions and different missing data scenarios. We also showed that a rather small latent space (four components) is enough to capture the essential of the input data distribution for missing data imputation.

Although FCAE consistently outperformed the baselines, the improvement was modest. Indeed, for the main experiment (H1), we reported an enhancement from 1.21% to 3.94% of the imputation error between FCAE and the better performing linear method. That being said, FCAE, being a deep-learned ANN, could benefit from being trained on larger databases.

To go further, this framework would benefit from handling categorical variables. Indeed, FCAE currently only handles continuous and ordinal variables (which are treated as continuous). Clinical expertise could also be beneficial to this work to penalize more clinically impactful errors (either for more important clinical scores or for some ordinal scores' successive categories). On another note, we think this approach could be valuable for longitudinal analyses of PD cohorts. An internal representation size of two could be used for visualization purposes, as it could project the data in a 2D map in which clusters of patients could be visually identified according to the evolution of their disease, allowing a better understanding of patient sub-typing. To complement this approach, homogeneous cohorts of PD patients having undergone different therapies could be used to visualize and understand the effects of the therapies better, allowing for counterfactual reasoning on retrospective patients.

In conclusion, this chapter developed and tested a powerful way of leveraging the crucial information residing in patient clinical tests by transforming it to a more readily usable form for downstream applications.

Chapter 3: ParDi

Another important source of information regarding the state of the disease of a PD patient is medical imaging. Our initial hypothesis is that MRI scans may carry out powerful information to characterize PD patients. We chose to focus on T1-MRI as this scan is consistently required to plan a DBS. It is mainly used for giving an anatomical support for future implantation and to detect important brain atrophies. As every medical scan, it is difficult to exploit because of its high dimensionality and myriad of information. To ease these difficulties, we decided to limit our analysis to specific structures. After an analysis of the state-of-the-art, we noticed that an informative way of analyzing a T1-

MRI for PD patients is by considering the bilateral striatum. In Chapter 3, we proposed a ML-based pipeline to extract informative biomarkers from the bilateral putamen and caudate nucleus (forming the striatum) from T1-MRI. This pipeline takes as input the patient's T1-MRI and an atlas segmentation of the structure of interest, giving as output a compressed vector of the shape displacement field of the patient's structure. We showed that this compressed vector is highly informative as a staging biomarker. Indeed, we managed to discriminate between several cohorts of PD patients and a control group, thanks to ML algorithms. Notably, we showed for the first time that the prodromal stage of the disease can be detected by analyzing the shape of the striatum.

On top of quantifying the ability of our pipeline to discriminate several PD cohorts, we also evaluated its capability of predicting motor symptomatology. It could be easily extended to other structures (notably structures visible on other scans than T1-MRI) and for other clinical scores. Concerning the latter, predicting the disparities of a clinical score for a single cohort could allow for addressing inter-population bias. Complementarily, doing a longitudinal analysis of striatal alteration would be an interesting follow-up, as it could allow a better understanding of the alteration dynamics of the striatum and its correlation with the evolution of clinical scores. Finally, this pipeline takes advantage of ML, but also relies on an image processing pipeline. We think a more data-driven oriented approach could be evaluated and compared to our method, for example by using CNNs on T1 and T2-MRI scans.

This chapter presents a promising way of extracting informative biomarkers from T1-MRI, and is complementary to FCAE in terms of representing the state of the disease of a patient.

Chapter 4: PassFlow

This collection of readily usable biomarkers allowed us to propose a pipeline, referred to as PassFlow and presented in Chapter 4, to predict the clinical outcomes of DBS for PD patients from input available during the screening phase. This phase mainly consists in selecting patients who could benefit from undergoing DBS. Clinician experts gather in clinical *staff* to discuss each patient, analyze the data, and, thanks to their expertise, anticipate the clinical effects that the stimulation could provoke, deciding accordingly on the best therapy. PassFlow tries to replicate this expertise by learning to predict clinical outcomes from retrospective databases. It consists in an SVM which takes as input:

— clinical information, using the encoder part of FCAE to compress it into a supervisedly-learned smaller latent space;

— morphological information, using the compressed shape displacement vector proposed in Chapter 3; and

— additional information, such as demographics.

The extensive work done in the previous chapters to transform rich information such as full clinical tests and T1-MRI into readily usable forms allowed PassFlow to predict significantly most of the post-operative clinical scores of the database. It outperformed a linear baseline and reached correlation coefficients up to 0.71, and was validated on a second database. This work is very novel as it is the first to use these many different data modalities as input. It is also the first which was tested across a large variety of different motor and non-motor clinical scores (21 clinical scores for each of the four post-operative phases). On top of being an interesting preliminary work for clinical use, PassFlow's surprising ability to predict most of the clinical outcomes of DBS demonstrates the importance of non-stimulation-related factors in anticipating the effects of the stimulation. To go further, it could hint on where to find factors explaining the great heterogeneity of DBS outcomes.

This work could be easily extended by taking additional biomarkers as input. Nonetheless, we showed in this chapter that on this database PassFlow did not have enough data to train completely. For this reason, increasing the dimensionality of the input could lead to worst results due to the curse of dimensionality.

A complementary work could be done by adding as input to PassFlow stimulation-related parameters, such as electrode contacts location and electrical parameters [4]–[6]. We think adding stimulation related parameters along with connectivity information could greatly enhance the predictive performance of our model. On top of allowing a comparison electrode location and electrical parameters prospectively, prior to the operation, it could also show how much of the clinical outcomes' variance can be explained by known and measured factors.

This chapter combines the two previous ones to propose a powerful data-driven pipeline which addresses a pressing clinical problem: how to determine if DBS could be beneficial to a patient, and if the operation is worth the risks that it involves.

Chapter 5: SepaConvNet

Chapter 5 is complementary to the previous ones as it addresses another clinical problem from the intra-operative phase: it proposes a DL-based way of locating the STN during the operation from MER spectrograms. To do so, we designed a custom CNN structure called SepaConvNet. This structure is based on 1D separable convolutions, allowing being a lightly parametrized ANN. We showed that SepaConvNet outperforms two state-of-the-art CNN structures (variants of ResNeXt), demonstrating its superiority for analyzing MER spectrograms. The results obtained with SepaConvNet, the first CNN tested on this problem, are comparable with, or even superior to state-of-the-art solutions based on feature engineering. Our method only requires 1 second of raw signal, while the state-of-the-art usually requires 10 seconds of pre-processed signal.

One limitation of this work is that it returns an independent prediction for each one-second window, therefore limiting its applicability in real time clinical condition. Indeed, for a difficult position to test, the system could make inconsistent predictions on successive windows and therefore be unusable by clinicians. To overcome this limitation, Martin *et al.* [195] extended SepaConvNet so that predictions are influenced by previous time steps. A Bayesian framework was tested as long as a trainable recurrent post-convolutions layer. It was shown that using an LSTM layer leads to a gain in accuracy on top of allowing system predictions to be more stable.

To be clinically usable, such a system must be transferable in other recording conditions. There are three possibilities regarding the expansion of SepaConvNet to new databases:

1. The training on one database can be straightforwardly usable in other recording conditions, providing that data pre-processing normalizes the data into the same space.
2. The training on one database can be usable to pretrain SepaConvNet, but the weights must be fine-tuned with data acquired with the second recording system. In this hypothesis, a small second database should be enough for good performance.
3. Knowledge acquired by learning on a database cannot be readily transferred to analyze a second database, therefore SepaConvNet must be retrained from scratch with a likely bigger database than for the second hypothesis.

Where traditional methods require additional engineering work in the form of feature-tuning (adapting, for example, the thresholds for spike detection) to be applied on new

recording conditions, data-driven methods only require a new round of data collection. This distinct advantage would be even more marked if one of the first two possibilities is verified.

In conclusion, this solution is encouraging as it shows that computer-assisted solutions, and more specifically DL could greatly enhance STN location during DBS. Indeed, this phase is crucial but also time-consuming and requires a lot of expertise. Our solution could speed up the STN location phase of DBS interventions, and therefore make the operation safer and more comfortable for the awake patient.

Section conclusion

This thesis is composed of four chapters and addresses two clinical problems encountered during the DBS workflow. Each chapter is a contribution that has been submitted to or accepted by an international journal or conference. First, a total of three contributions design a CDSS able to significantly estimate most of the clinical outcomes for PD patients undergoing DBS. First, a system was proposed to synthesize and curate pre-operative clinical information. A second one extracts informative biomarkers from T1-MRI. The last contribution combines the two prior systems to predict clinical outcomes from the pool of curated biomarkers. Another CDSS locates the STN from MER acquired during the surgery, only requiring one second of signal. Each contribution raises interesting discussion points, such as limitations and avenues for future work.

6.2 Discussion regarding the introduction's hypotheses

In the introductory chapter of this thesis, we set two hypotheses:

— Pure data-driven methods can be relevant to solve clinical problems on the DBS workflow. Not relying on aggressive feature selection methods or feature engineering could lead to new levels of performance and address more ambitious purposes.

— The best models to support such a pure data-driven paradigm are deeply-learned ANNs. Indeed, ANNs have proven themselves to be powerful in many applications and their computational power allows for complex links between inputs and outputs.

The four contributions proposed in this thesis offered additional answers regarding these hypotheses, in the context of DBS for PD:

Are pure data-driven approaches relevant for assisting clinicians in DBS?

We used data-driven approaches for all four contributions. Here are the conclusions relatively to this question for each of them:

— Chapter 2 - PatiNAE: while no comparisons are made with non-data-driven approaches in this chapter, we showed that pure data-driven algorithms (linear and non-linear) can successfully understand the underlying distribution of input clinical data, and that this property can be exploited for both data compression and imputation without requiring explicit knowledge of the field.

— Chapter 3 - ParDi: in this chapter, we took advantage of a machine-learning-based pipeline to use the whole shape displacement fields of four subcortical structures as staging biomarkers. The results are a strong argument in favor of data-driven methods to find such biomarkers. Indeed, our method considers the entire shape alteration simultaneously instead of simplifying the alterations of subcortical structures (by taking for example only its volume or its thickness), or by doing a mesh point-wise analysis, like most of the studies in the literature. Employing this data-driven paradigm obtained novel results in this research field: we showed for the first time that striatal alterations can be used to detect prodromal phase of the disease. In this chapter, the only clinical hypothesis we exploited was the selection of the structures to limit its scope. Nonetheless, we think this method could benefit from being replicated for other structures and on other imaging modalities, without any clinical hypotheses, maybe allowing discovering unsuspected links between neuroanatomy and symptomatology.

— Chapter 4 - PassFlow: in this chapter, we used the methods and biomarkers presented in Chapters 2 and 3 to predict 21 clinical scores at four different phases after the surgery. Here, the advantages of pure data-driven methods have been exhibited too, for two reasons. First, the great span of information used by the predictive model allowed to predict most of the clinical scores significantly, with overall good

performance. Second, the validated pipeline does not rely on clinical hypotheses and manages to use a great pool of input information, and it can therefore be easily extended and adapted to other use cases. We took advantage of this property to test our model on 84 scores for the main database, but also 20 different scores on another database. To the same extent, other inputs can easily be added to the model and their relevance can be quickly identified. In other words, on top of having interesting performance, the adaptability and applicability of pure data-driven methods have been highlighted in this chapter.

— Chapter 5 - SepaConvNet: this chapter consisted of using a deeply-learned CNN to identify the STN in MER. Here, yet another interesting property of data-driven approaches have been shown. Indeed, for this problem, common state-of-the-art methods are based on feature engineering, requiring strong hypotheses and years of tests and expertise to be effective. With our method, we managed to achieve comparable performance from the whole, raw MER information. DL managed here to converge to a comparable, or superior level of performance that required a decade of research with more traditional methods. Additionally, we believe our method is more easily applicable to other hospitals and recording hardware.

Through four distinct contributions, this thesis successfully highlighted three distinct advantages of data-driven methods over feature-based ones:

— They can achieve greater results than regular methods, potentially exploiting additional information from the data not covered by commonly used hypotheses.

— The fact that they don't rely on such hypotheses allow to obtain satisfying results in a short amount of time, compared to methods that could require years, or decades of expertise.

— Finally, once data-driven methods have been developed and validated for one use case, they can easily be extended or shifted to new problems or conditions, such as new cohorts or new data-acquisition pipelines.

Nonetheless, the problem of the curse of dimensionality along with the limited sizes of databases in DBS still causes performance and convergence issues. For this reason, the employed methods have to be designed and adapted to control this problem's impacts.

Are artificial neural networks relevant in the perspective of employing a pure data-driven paradigm?

We tested ANNs for various tasks on each contribution of this thesis, with the hypothesis that their computational power could have better performance than other models, or that their customizable nature could address specific issues of each problem.

— Chapter 2 - PatiNAE: in this chapter, we directly compared the performance obtained with an ANN (a homemade deeply learned AE) versus two linear baselines. We hypothesized that AE could perform better than PCAs thanks to the activation functions of neurons allowing for complex non-linear combination of the input data. This ability of deeply learned AEs to compute complex functions which represent highly non-linear input distributions could allow obtaining a more optimized latent space. We trained our AE with a problem-tailored loss function to explicitly penalize errors in missing data imputation, and thus obtained better performance for data imputation with our method. While significant, this improvement was modest and necessitated many developments on the ANN topology: simpler AEs tended to perform worse than linear baselines. We had to test several architectures and run many tests and HPOs to find a topology that was able to take advantage of the non-linearities of the data to unlock better performance. This illustrates the difficulty of training deep-learned ANN successfully. Indeed, ANNs are multi-parametric, requiring a lot of optimization on the choice of architecture and hyper-parameters. In addition, heavy and powerful architectures can greatly suffer from overfitting on small, heterogeneous databases.

— Chapter 3 - ParDi: this chapter is the only one where we didn't use any ANNs, despite many attempts on two parts of the pipeline. First, concerning striatal shape compression, we unsuccessfully tried various deep-learned AE. At best, AEs could perform similarly to PCA at the cost of much greater training time. Second, concerning the classification and regression tasks, several topologies of ANN have been tested, but all of them performed significantly worse than SVM and RF and were thus not retained. We think in both cases the problem was the limited database size: for compression, shape displacement vectors' sizes were an order of magnitude greater than the number of training points. This compression task being unsupervised, gathering more T1-MRIs from other databases to compute more shape displacement vectors could have allowed AE to perform better by substantially in-

creasing the size of the training database. Nonetheless, the image processing pipeline being computationally heavy, we chose not to retain this option.

— Chapter 4 - PassFlow: the usage of an ANN in this chapter was similar to the one of Chapter 2. Indeed, we reused the same ANN topology to compress clinical data while being robust to missing values. In contrast to Chapter 2, we did not train the encoder unsupervisedly thanks to a decoder, but supervisedly with a sigmoid activated output neuron in charge of the regression task. Unexpectedly, this second strategy was more efficient. Again, database sizes can be the explanation: in this chapter the ratio between input dimensionality and database's number of samples was smaller than for Chapter 2 (26.7 versus 4.7). Therefore, it is possible that an unsupervised learning strategy fails to capture the underlying distribution of input data and capture a larger amount of noise instead. The supervised strategy, on the other hand, could have a good first approximation of a relevant latent space by simply connecting the pre-operative scores of the post-operative clinical score to predict to the internal representation.

— Chapter 5 - SepaConvNet: for this Chapter, we used CNNs to analyze spectrograms using 1D convolutions. We obtained satisfactory results thanks to the power of CNNs, arguably the most common and powerful type of deeply-learned ANN. Contrarily to the previous chapters, here, there is no direct alternative to CNNs for image analysis: other methods use a set of features or a projection of the image in a smaller space, but not the raw images directly. Therefore, for this application, using a data-driven, not feature engineering-based, method is synonymous with using a CNN. Our proposed CNN, outperforming other CNNs, was light in parameterization (around 16 thousand parameters, versus, for example, 60 million for the winning CNN in the ILSVRC 2012 challenge). The fact that heavier versions of this structure were not better performing (as the HPO also covered heavier versions of SepaConvNet) indicates that this additional available computational power was not exploited, i.e. there were no benefits of allowing for more features to be computed, or that there were no interesting higher-level features learned. Yet again, one of the reasons could be the size of the databases (in the order of tens of thousands of images, versus more than 15 million, for the ImageNet database used in the ILSVRC 2012 challenge). Moreover, these images originate from 57 patients only, and we expect a high inter-subject variability on MER analysis. In this case, even more data can be necessary to detect and learn only patient-invariant features from raw

spectrograms and learn to ignore the large, but uninformative, variance between patients.

To summarize, ANN stood out in two applications. First, deep-learned AEs have been applied to a compression problem of unstructured data by training it with a custom, problem-tailored loss function. On top of the theoretical superior computation power of ANNs compared to linear method, their flexibility allowed to address a missing data problem specific to clinical tests straightforwardly. Second, a CNN was able to classify complex structured data thanks to 1D separable convolution kernels. The ability of ANNs to be easily adapted according to the specificities of the input data increases their relevance, allowing to use a pure-data-driven paradigm for the problem. On the two other applications, ANNs did not achieve better performance than simple methods. For both applications, we did not identify input data or problem specificities that could be addressed by adapting the ANN's topology or training method. Finally, across each application, small database sizes were an identified limiting factor.

Taken together, it indicates that ANNs should be considered as a *means* and not an *end*: using deep learning shouldn't be an objective on its own, but should rather be considered as a powerful tool in some specific, identified use cases. Applications where ANN could shine can be identified by analyzing the problem and the data. Providing that databases are large enough, their flexibility and their computational power can be exploited to outperform more static and less scalable methods. Advanced techniques, such as transfer learning [196] or multitask learning [197] can also be employed with ANNs providing that complementary and relative databases are available. Besides that, standard ANNs organized as MLPs have no reasons to be employed on a particular ML or data-driven problem and should not be considered as fundamentally superior to other ML models.

Section conclusion

This thesis illustrated three advantages of data-driven methods: the non-reliance on pre-existing hypotheses and knowledge can be profitable in terms of performance, and also mitigates the large amount of time required to test and validate clinical hypotheses or to design features that are a translation of these hypotheses. Data-driven methods are also quickly transposed to other use cases, because the methods can be reapplied straightforwardly. While the advantages of data-driven methods are clear, the conclusion is less so for deep learning. Even though AE have been successfully used, the performance improvement was modest. For spectrogram classification, this improved performance due to CNNs was more convincing, likely due to them being a relatively well understood and widely used tool for purely data-driven signal processing.

6.3 Limitations and perspectives

In this section, we will raise current limitations of data-driven approaches and ML for healthcare research. Following that, we will present some interesting application perspectives for such methods.

6.3.1 Limits and future works

Inconsistent validation methods and databases

Methodological research in ML relies a lot on 1) benchmarking proposed methods versus the state of the art and 2) reproducing the methods on other databases and use-cases. Unfortunately, both of these points are problematic in medical research. Indeed, benchmarking methods either consist on using the same database and validation method as the baseline's paper, or by implementing the baseline method to evaluate it on other validation conditions or other database.

When it comes to data-sharing, several issues can be encountered in medical research, in addition to ethical issues and data privacy concerns. First, the clinicians and labora-

tories themselves tend to be reluctant to share their clinical data. Indeed, clinical data gathering implies a lot of manual work to retrieve and curate the data in the case of retrospective studies and can perturb the clinical routine in case of prospective studies. In order for the data to be readily usable for research, a lot of engineering work is required and the data collection can take up to several years. For example, the Rennes DBS database consists of more than eight years of data collection and required the work of several clinicians and engineers to be usable by data-scientists in the MediCIS team. Considering this extensive amount of work, keeping the database private in order to have enough time to exploit it and publish with it is a legitimate strategy. Additionally, the clinical routine is highly variable between different hospitals, but also within hospitals along the years, with the natural evolution of surgical techniques. This heterogeneity multiplies the number of possible studies, but also complicates data merging between different databases. Multicentric databases are rare on DBS for PD, as deciding on a common clinical workflow and collecting data according to a common protocol requires a lot of time, human resources, will and involvement from every partner. The lack of standards and tools to record or share data, resulted in scattered and heterogeneous DBS datasets [198]. This data sharing (or data mutualisation) problem makes the results comparison of methods developed by different teams almost impossible.

An additional problem is the inconsistency of validation techniques used in the literature. Baxter *et al.* [199] reported the validation techniques employed in the papers retained in the systematic review done in Section 1.3. Results are presented in Figure 6.1 and denote a great heterogeneity of the techniques used.

There is also no consistency in the validation techniques employed on individual research topics, preventing from comparing the results. A good illustration is the recent work done by Khosravi *et al.* [200] on STN localization with MER. This work is relatable to our SepaConvNet study, but no direct comparison of the reported results could be done without re-implementing their method for several reasons:

— While using the same database, their work uses a more complete version of it (100 patients instead of 57 used in Chapter 5);

— The evaluation method is not the same and information allowing the exact same partition is not provided; and

— They used accuracy as a comparative metric, whereas we used balanced accuracy.

For these reasons, even if the performance reported by Khosravi [200] *et al.* seem greater than SepaConvNet (92% of accuracy reported on the training set), we found

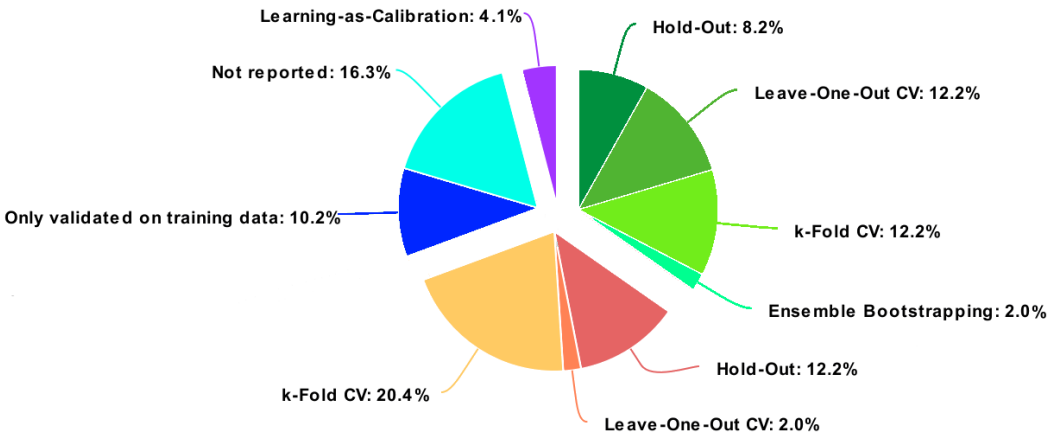


Figure 6.1 – Validation techniques of the papers covered in the systematic review of the literature on ML methods in DBS presented in Section 1.3. Papers with proper patient-wise data splitting are in green, papers with identified data leakage are in red, and papers for which it was not specified are in blue. Source: [199].

worst results after having replicated their method using our evaluation conditions (70.7% of balanced accuracy against for 80.5% for SepaConvNet on the testing datasets using cross-validation).

On another note, biases or lack of information regarding the evaluation methods can prevent from having comparative baseline's results. The review of Baxter *et al.* [199] illustrates this point as only about 35% of the articles (in green in Figure 6.1) employed a proper patient-wise splitting, which minimizes the opportunity for data leakage between training and testing data. For example, Sivaranjini *et al.* [201] proposed a CNN to diagnose PD subjects from controls from a single slice of T2-MRI. They split the data into a training set and a testing set, but they never specified if the data splitting was done patient-wise. Not performing patient-wise splitting allows for slices of a single patient's image to be present in the training set and the testing set, the reported performance should be greatly over-estimated because of the algorithm in training time may have access to information about the particularities of said patient and would not be representative of prospective use. Without being specified, it is unfortunately not possible to safely compare the performance of any method to their reported one.

The other way of comparing methods, if validation conditions cannot be replicated, is to redevelop pieces of software in order to run the pipeline in common validation conditions. Unfortunately, method reproducibility is also an issue in biomedical engineering.

Contrary to other engineering research fields, software and scripts developed by research teams are rarely shared publicly. For example, in the systematic literature review presented in Section 1.3, none of the papers shared their code to assist with reproducibility and therefore comparative studies.

Sharing code also limits the possibility of development errors, such as data leakage, which can be dramatic if undetected. Vandewiele *et al.* [202] reproduced the methods and invalidated the results of 11 studies claiming to have nearly perfect accuracy (up to 99.4%) for premature delivery detection with electrohysterogram records. They found a common mistake in data partitioning by over-sampling the data before splitting it in training and validation sets. Corrected results fall to, at best, an accuracy of 60.85%. The first study suffering from this bias was published in 2013 [203]. It was not clearly mentioned in their paper in which order oversampling and data splitting was done, and it necessitated seven years and a whole new round of software development to reproduce the method and check for the validation bias. In this case, publicly shared code could have allowed direct code review by the community and maybe would have avoided ten more flawed studies on the six following years.

In this thesis, we made public, so far, part of the code developed for Chapters 2 [1] and 3 [2].

Data collection and sharing

It is common knowledge that training ANN requires an extensive amount of annotated data, and contributions of this thesis confirm this point.

On several occasions, experiments indicate that database sizes may be too limited yet to support the utilization of deep learning. It also seems to indicate that, in the problems studied by this thesis, the success of data-driven methods is more a matter of data collection than a matter of model computational power. In other words, a simple method could perform well on a large database, where a powerful model cannot compensate for small database size. While good results can be achieved by handling data appropriately, performance still stagnates until extensive databases are gathered. Complementarily, the proliferation of DBS-related databases could allow for *transfer-learning*. Transfer-learning is a technique which consists of pre-training an ANN on a related problem or database to initialize the ANN's trainable parameters, before retraining it on the actual problem and

1. `https://github.com/m-prl/PatiNAE`
2. `https://github.com/m-prl/ParDi`

	PPMI	Rennes	PrediSTIM
Cohort	Prodromal PD, Early PD	DBS PD (STN, GPi, VIM)	DBS PD (STN)
Imaging scans	T1 MRI/T2 MRI/ fMRI/DTI/PEC/ SPECT/CT	T1 MRI/T2 MRI/ CT (perop, postop)	
Clinical scores	MDS-UPDRS (all parts), Physical Activity Scale for the Elderly, Modified Schwab & England ADL, SCOPA-AUT, Clinical Cognitive Categorisation, Geriatric Depression Scale (Short), QUIP, State-Trait Anxiety Inventory, Benton Judgement of Line Orientation, Hopkins Verbal Learning Test, Letter-Number Sequencing (PD), MoCA, Semantic Fluency, Symbol Digit Modalities Text, Epworth Sleepiness Scale, Features of REM Behaviour Disorder	AES, AMDP-AT scale, Ekman 60-faces test, Categorical Fluences, Lexical Fluences, Verbal Fluences, Hoehn and Yahr, MADRS, MDRS, Schwab and England, Stroop, TMT A, B, B-A, UPDRS (all parts)	Non-motor fluctuations evaluation, Miami hallucination scale (total, visual, auditive, somesthesic, gustatory, olfactive), PDQ39, CGIS, SCOPA-PS, MDS-UPDRS (all parts), Hoehn and Yahr, Schwab and England
Phases	Preop	Preop (-6M, -3M), postop (+3M, +6M, +1Y, +3Y, +5Y)	Preop, postop (+1Y, +3Y, +5Y)
Stim. params.	X	Yes	Yes
Electrophy. signals	X	Yes	No

Table 6.1 – Data modalities collected for the three databases used in this thesis.

database. This technique is regularly used for CNN architectures in computer-vision and was proven powerful many times, even in medical image analysis [204].

Unfortunately, for DBS, extensive and uniform databases are complicated to gather, as data modalities are inconsistent throughout time and across centers. While this heterogeneity allows for a wider range of research topics and methods, it also prevents merging retrospective databases as they don't share a great number of common data modalities (especially when it comes to clinical tests). Additionally, clinical workflow can evolve at a clinical site and the data modalities acquired can therefore change, which can disrupt uniform data collection. Table 6.1 illustrates this by showing the data modalities collected on the three databases used on this thesis.

Finally, data mutualisation is greatly affected by the sensitive and private nature of personal health records. Gathering personal medical data for research involves many legal constraints and must be compliant with various regulations, such as the General Data Protection Regulation (GDPR) in the Europe Union or the Health Insurance Portability and Accountability Act (HIPAA) in the USA, which is a great asset for human rights and privacy, but also complicates data sharing and slows the retrospective or prospective creation of medical datasets. For example, this thesis involved the review and acceptance of four different data collection and data utilization protocols by local research ethic committees.

Towards collaboration, standardisation and community-powered initiatives

The future of biomedical research and its integration in medical workflow depend on the ability of various clinical and research actors to embrace common practices and rigorous standards, especially for data-driven methods [205].

Toward this objective, some libraries have been proposed in order to facilitate uniform software development for DBS, therefore facilitating code sharing and reproducibility by ensuring a common software back-end and application programming interface. For medical image processing, *ITK* [206] and *Slicer* [207] are the two most commonly used libraries, proposing functionalities such as segmentation or image denoising. Both of these libraries rely heavily on their community to maintain, review and document the code. For DBS itself, *Lead DBS* [208] offers many functionalities that promote research, such as electrode localization, connectomes, and multiple non-linearly registered atlases. Regularly, updates are proposed to integrate the works of researchers communities. Concerning DL, alternatively to generic frameworks (such as *Keras*, *TensorFlow* or *PyTorch*), *NiftyNet* [209]

offers functionalities to ease development and utilization of deep neural networks for medical imaging analysis specifically. Additionally, a model zoo is included in the framework, allowing to perform common tasks (such as regression, segmentation or image generation) quickly. Finally, *NVIDIA Clara Imaging*[3] offers tools to ease medical imaging annotation, as long as APIs to train deep learning models and integrate them in clinical workflows.

Another crucial aspect is data standardisation. Indeed, it is well known that the *Digital Imaging and Communications in Medicine (DICOM)* format lacks standardization, which results in many difficulties in replicating a method with another dataset. In the purpose of addressing this issue, the *Brain Imaging Data Structure (BIDS)* [210] format was proposed as a standard to organize and name medical images. Complementarily, an online tool[4] was proposed to validate if a dataset is compliant with the BIDS specifications.

Finally, community challenges have an important role in collaborative research as it offers a framework for scientists to work on common problems and/or datasets with various methods. Some of the libraries mentioned in this section offer regular workshops and hackatons (such as *Slicer* [211] or *Lead-DBS*). Complementarily, there exist platforms (such as *Grand Challenge*[5] or *Kaggle*[6]), that share datasets and propose public contests on various biomedical research problems.

We have the strong belief that such approaches are at the heart of healthcare research, as they speed-up recurring software development, reduce software errors and bugs as the code is reviewed by multiple persons, and allow for community emulation and federation, and multi-actors collaboration.

Hardware requirement and carbon footprint

The performance of ML and specifically DL come at the cost of an extensive amount of computational time and resources. Indeed, on top of being usually heavyweight, ANNs require topology and training hyper-parameters tuning, necessitating hundreds or thousands of neural networks training for a typical research paper.

To speed-up ANN training, one or several powerful Graphics Processing Units (GPU) are increasingly mandatory. In research laboratories and industries, computational clusters are often used in order to centralize the computational resources, avoid to leave a GPU

3. https://developer.nvidia.com/clara-medical-imaging
4. `https://bids-standard.github.io/bids-validator/`
5. `https://grand-challenge.org`
6. `https://www.kaggle.com/competitions`

Emissions of	kgCO2e	Paris-New York round trips
BERT training	658	1
So *et al.* [213]	284019	434
My thesis	4247	6.5

Table 6.2 – Carbon footprint of the electricity consumed during my thesis, compared with the training of BERT (a state of the art model) and the total estimated emissions of the breakthrough of So *et al.* [213] on English-to-German translation. To give an order of magnitude, the equivalent amount of CO2 equivalent emissions of a Paris-New York round trip by plane for one passenger is given in each case.

unused and therefore to limit the costs. To go further, external computation clouds are also greatly used for even more flexibility.

On top of financial considerations, this heavy hardware requirement also has a large carbon footprint, both for hardware production and transport, and hardware energy consumption during training and evaluation.

Only a few papers exist in the literature which investigate this. Among them, we can mention the work of Strubell *et al.* [212] which quantified the carbon footprint of the development of DL models in NLP. They notably reported that a simple training of BERT, a state of the art NLP model, emit around 658 Kilograms of carbon dioxide equivalent (kgCO2e) which represents the CO2 emissions of a Paris-New York round trip by plane for one passenger (654.6 kgCO2e with the ICAO calculator [7]). They also reported that the new state-of-the-art performance on English-to-German translation achieved by So *et al.* [213] came at the cost of 284019 kgCO2e (434 Paris-New York round trips by plane for one passenger), for a modest improvement (0.34% of improvement on the comparison metric versus the better performing baseline). We evaluated the energy consumption of this thesis at 9.5 MWh. Given the average European emission rate of 0.447 kgCO2e per electrical kWh consumed [214], the thesis emitted around 4247 kgCO2e, representing 6.5 Paris-New York round trips by plane for one passenger. All of these numbers are recapitulated in Table 6.2.

Considering the short and long term dangers of climate warming, to us, the trade-off between expected concrete interest of any ML research contribution versus its expected greenhouse emissions should be better considered, even in medical research. Toward this idea, Lacoste *et al.* [215] proposed an online machine learning emissions calculator [8], as

7. `https://www.icao.int/environmental-protection/Carbonoffset`
8. `https://mlco2.github.io/impact/`

well as recommendations for limiting the carbon footprint of ML applications. Henderson *et al.* [216] proposed a framework to calculate real-time energy consumption and carbon emissions of ML experiments, supporting their ambition of promoting systematic and accurate reporting of energy and carbon usage in ML research papers.

6.3.2 Usability of ML methods as CDSS

Non-knowledge-based CDSS are data-driven tools designed to assist clinicians for decision-making tasks. This notion is central in this thesis as the four contributions allowed to make two distinct CDSS-core predictive systems. In order to be used in clinic, any healthcare technology must be rigorously assessed, which is the subject of Health Care Technology Assessment (HCTA), a multi-disciplinary field concerned with the evaluation of the technology's short-term and long-term consequences [217]. Concerning the assessment of systems for image-guided interventions, Jannin *et al.* [218] distinguished six levels of assessment, from the *technical feasibility and behavior* of a system to its *social, legal and ethical impacts*.

The particular case of ML-based predictive systems involves specific complementary factors including: performance, resource usage, reliability, representativity, ethics, and acceptability.

Performance

It is hard to define a quantitative threshold above which performance of a ML method could be considered as sufficient due to clinical workflow considerations. At first glance, outperforming the clinician may seem mandatory, especially given the high performance of recent computer vision algorithms trained on large-scale databases. However, this is not always feasible or even measurable. For example: SepaConvNet proposed in Chapter 5 cannot theoretically outperform clinician annotation as these annotations *are used as* the ground-truth.

Resource usage

Another interest could be the speed-up offered by the usage of an automatic tool in terms of either the speed of processing, or the speed at which the necessary data can be acquired. The later scenario is demonstrated again by SepaConvnet; the system is able to consistently give a prediction in a second, which is not possible for an expert. Any speed

advantage comes with trade-offs in terms of cost and the availability of specific hardware. As powerful algorithms in ML are known to be GPU, Central Processing Unit (CPU) and memory intensive, especially for deep-learned ANNs, speed could be gained by clinical investment into computing power, but this is limited by the budgetary restrictions of the clinic in question as well as what computational equipment is currently available in clinic. This concern motivated us to make SepaConvNet as light as possible, in order to guarantee its compatibility with most computers used in operating rooms.

Reliability

The reliability of a predictive system could be defined as its ability to make consistent predictions: changes in prediction are made only in response to meaningful changes in input. Unfortunately, it is well known by the ML community that this property is not necessarily fulfilled even if a model exhibits very good retrospective performance. A good illustration is adversarial examples, which consist in tricking a predictive system into drastically changing its prediction by almost imperceptibly changing the input. This phenomenon has been spectacularly demonstrated by Ren *et al.* [219] who conducted an extensive study on this domain, showing an example where an imperceptible perturbation on an image of a panda causes the CNN to predict it as a gibbon with 99.3% confidence (despite its initial prediction on the unperturbed image being correct).

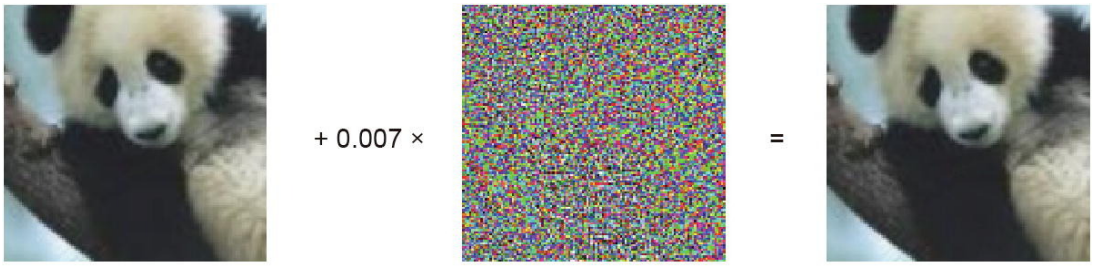


Figure 6.2 – Illustration of an adversarial attack: when adding to an image a noise imperceptible by the human, the machine can drastically change its prediction. Source: [219].

This problem is crucial in clinical application where errors could have dramatic consequences. For example, PassFlow takes as input clinical testing of a patient. It is well known that some items of clinical tests and questionnaires can be noisy, either because

the clinical state of a patient is not consistent across time, or because the assessment is not consistent between clinicians or, in case of self-rating questionnaires, patients. In this example, a difference of one point in, for example, the UPDRS3 scale shouldn't greatly impact the predictions of the system.

Representativity

Another problem arises from the distribution of the databases used to train ML models and whether or not that distribution is sufficiently representative to reflect the variety of conditions and anomalies seen in clinic. Indeed, if a subgroup of patient is over-represented in the retrospective database used to train the model, predictions may be biased towards this group at the expense of other under-represented groups [220]. Moreover, if a non-typical patient is encountered prospectively, falling outside of the training database distribution, the predictions of the system can be considerably worsened, or at worst become completely random.

For these reasons, an interesting future complementary work would be to bundle the predictive system with a method capable of accurately estimating the confidence of the predictions. The system should be aware of if a prediction is risky or not, and this information should be easily readable by the users. Several preliminary attempts have made to give PassFlow this capability, but they did not give interesting results.

Ethics

Another critical aspect of ML-based prediction in CDSS are its ethical implications. According to Rigby [221], "current policy and ethical guidelines for AI technology are lagging behind the progress AI has made in the health care field". On top of privacy and confidentiality issues, a major concern lays in the legal liability of medical malpractices caused by AI errors. Sullivan *et al.* [222] said that "if a patient becomes injured by [the] use of an AI technology (black-box AI in particular), current legal models are insufficient to address the realities of these innovations". There is no doubt that a more solid legal framework should be proposed in order integrate AI in clinical routine. Finally, the agreement of patients is mandatory in the usage of AI technologies. Fenech *et al.* [223] conducted a survey on adults in United Kingdom outlining that 63% of respondents think AI shouldn't be used replace practitioners on tasks they usually perform, such as suggesting treatments.

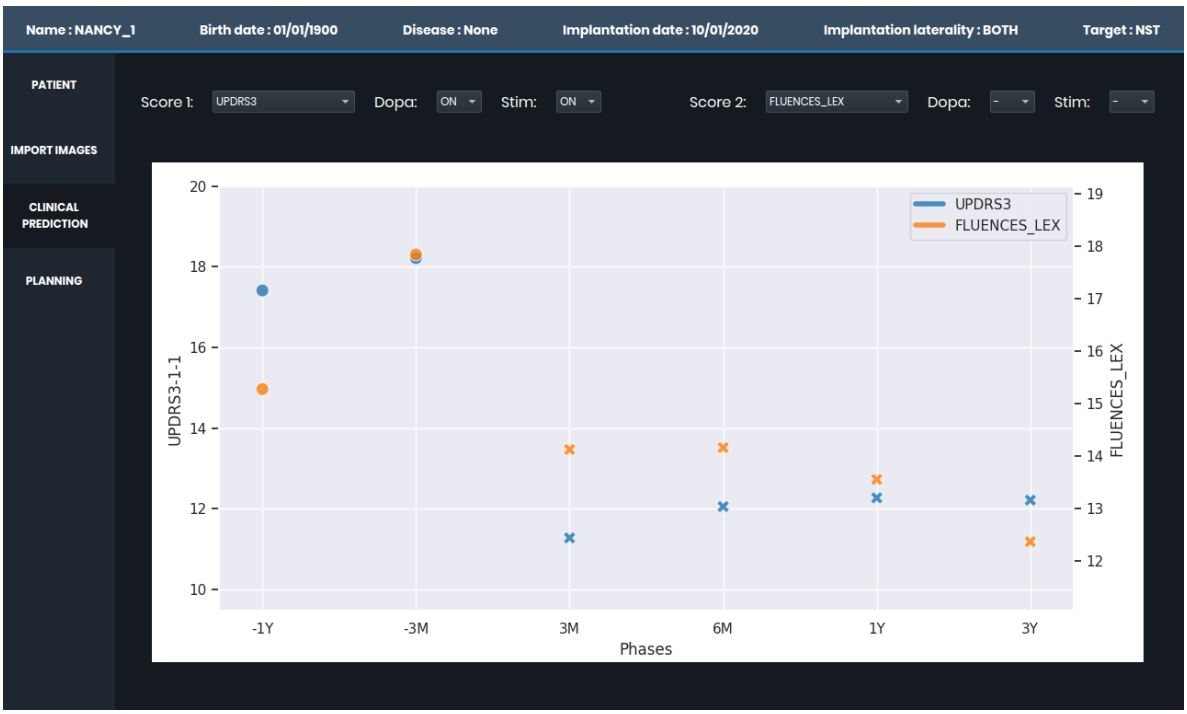


Figure 6.3 – PassFlow has been integrated as a plugin into PyDBS. After selection of a patient with known pre-operative information, the plugin shows one or two selected scores at different time-points. Pre-operative scores originate from the database (or are imputed by our auto-encoder presented in Chapter 2 if missing), and post-operative scores are predicted by PassFlow.

Acceptability

Even if the previous issues are solved, a ML-based CDSS could be integrated and used in the clinical routine only if the clinical staff accepts, values, and trusts it. In order to extensively study this crucial point, we conducted an acceptability study for PassFlow [224]. This study consisted in a questionnaire based on the Unified Theory of Acceptance of Technology (UTAUT) model, in two parts. Twenty-five French DBS practitioners participated in this study (11 neurosurgeons and 12 neurologists, neuropsychologists and one radiologist). A demonstration video of the PassFlow system integrated in the PyDBS software (showed in Figure 6.3) was presented to the clinicians before they answered the first part of the test. Then, they interacted with the system via video-call, seeing post-operative predictions for three patients (one with positive DBS outcome, one with mixed outcomes, one with negative outcomes) before answering the second part of the test.

Our acceptability study found some interesting preliminary results including:

— Results are overall consistent between the neurosurgeon group and the neurologist/neuropsychologist group.

— Practitioners would value such a tool for assistance for complex cases. The results are more mixed when it comes to accelerating simple cases (Figure 6.4a).

— They would have changed the chosen therapy for two patients (the ones showing negative or mixed post-operative clinical outcomes) if post-operative outcomes were known beforehand. This confirms that, providing accurate enough predictions, such a tool would enhance the care of the patients by reducing errors.

— Almost unanimously, the black-box effect of the prediction system was considered as a problem. Practitioners would like to understand the reasoning of the system before they use it.

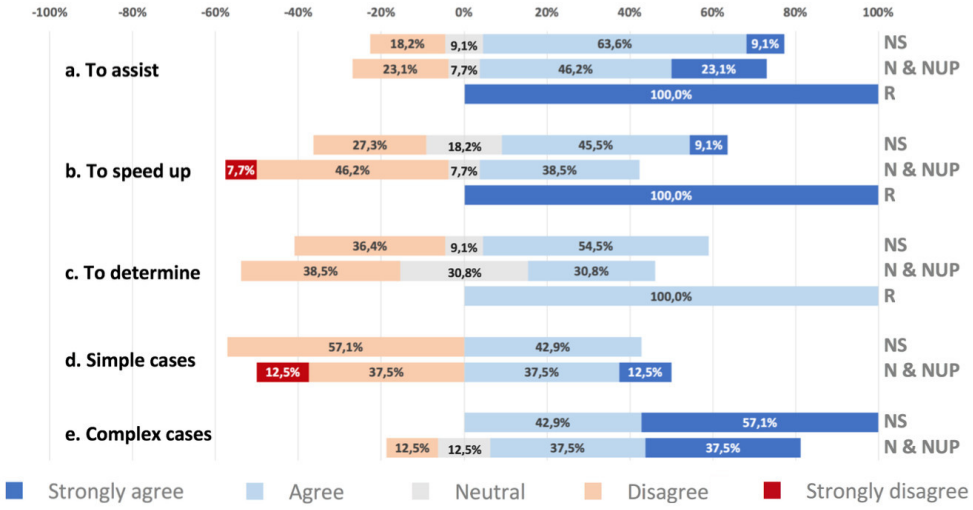
— All participants expressed a desire to use PassFlow within the next year if it were to become available (Figure 6.4b).

6.3.3 Perspectives

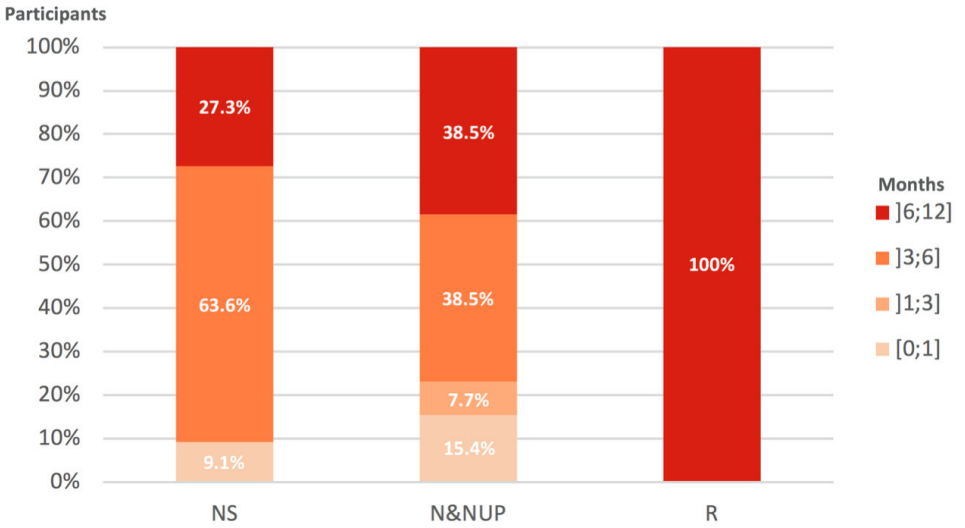
On top of clinical applications addressed in this thesis, we identified some other applications where data-driven methods could be appropriate and beneficial.

First, post-operative cartography could greatly benefit from data-driven methods. Indeed, the research space of this problem is huge as a lot of combinations of different parameters (contact(s) activated, voltage, frequency, pulse-width) can be considered [225]. Only a small number of these combinations can be tested considering the time necessitated, along with the discomfort for the patient. This would be exacerbated with the use of multi-directional leads, which are considered as an interesting innovation for DBS [226]. We think a data-driven method could be beneficial here, as it could virtually test a great number of combinations and return the best ones. Moreover, such a data-driven method could take into account the particularities of the patient in the form of bio-markers, connectivity and diffusion information.

Second, as stated in the systematic review in Section 1.3, an extensively studied problem is LFP analysis for closed-loop stimulation which adapts to the disease fluctuation instead of delivering stimulation continuously. This paradigm is often considered as a major opportunity to improve DBS [226], [227], and we think SepaConvNet could be helpful in such a closed-loop system. To go further, we could even imagine the design of a system



(a) Main utility in the decision-making process.



(b) Behavioral intention to use (desire to use the system in the next N months if made publicly available).

Figure 6.4 – Two of the main results highlighted by Diot *et al.* [224]. NS stands for neurosurgeon. N&NUP stands for neurologist and neuropsychologist, R stands for radiologist. Source: [224]

which could dynamically change the stimulation parameters according to the analysis of the LFP.

Third, we think MER data acquired during the operation could serve more ambitious purposes than merely locate the target. Indeed, we think that MER analysis could predict post-operative clinical outcomes resulting from stimulation at several depths, and therefore could predict the best stimulation target point and parameters. For such a hypothesis, a powerful method, such as CNN would be likely necessary.

Fourth, the area where DL shines the most is image analysis. We think further work could be done towards the analysis of multiple imaging modality simultaneously, in order to find cross-modality patterns in cortical or sub-cortical structures. A lot of problems could be addressed with such a method, from pre-oeprative planning to patient clustering. For example, a pipeline able to analyze T1-MRI and T2-MRI simultaneously could allow the identification of most of the structures of interest for pre-operative target and trajectory selection, allowing not to rely on atlases and not to suffer from patient-to-atlas registration errors.

6.4 Conclusions

In this thesis, we proposed four different ML tools. The first three tools address problems of patient selection and pre-surgical planning for DBS, including an ANN to curate clinical information, a pipeline to extract bio-markers from the striatum on T1-MRI, and another pipeline that combines these information to predict various clinical outcome of DBS. The final tool was a custom CNN structure able to accurately and quickly localize the STN from MER during a DBS intervention. Both of these applications are answering to precise and crucial clinical needs, gave interesting results, open new perspectives and are subject to enhancements. Useful CDSSs could be developed to assist practitioners and enhance the safety and effectiveness of surgeries. On top of that, clinical and neurological knowledge could be unveiled with methods that do not necessitate pre-existing hypothesis. The potential and interest of a pure data-driven research paradigm have been illustrated throughout this thesis, in which the flexibility and computational power of deep learning, as well as traditional machine learning, were key elements. There is no doubt that there are multiple impactful and crucial problems which could be addressed using such approaches.

Nonetheless, while a number of methodological improvements remain to be discovered, a ceiling currently exists regarding the performance and applicability of such methods in clinical routine until several limitations are overcome. Firstly, there exists a data collection and sharing problem, preventing both from gathering big enough databases for ML to really shine, and to effectively compare methodological work. The later issue is amplified by the lack of parsimony in evaluation methods, and the reluctance of data scientists to share their code.

Complementarily, larger obstacles remain to be addressed in order to expand the applicability of AI in neurosurgical care. A first technical issue is the reliability of ML. In order to be applied despite their black-box nature, ML models need to not suffer from biases in the training databases, and also be provably robust to variations in the data similar to adversarial attacks. To go further, a legal framework has to be developed around the usage of ML-based CDSS. The acceptability of such methods for both medical practitioners and patients remains a crucial and unavoidable question.

It is undeniable that ML, DL and data-driven approaches have a promising future in DBS, and that machine and human intelligence have many ways to collaborate.

List of Figures

List of Tables

Glossary

AAM Active Appearance Model p. 91

AE Auto-Encoder pp. 41, 66–70, 77, 149, 157, 159

AI Artificial Intelligence pp. 32, 38, 41, 170, 175

ANN Artificial Neural Network pp. 34, 39, 50, 54, 67, 69, 117, 122, 123, 149, 150, 152, 154, 156–159, 163, 166, 168, 174

AUC Area Under the ROC Curve p. 52

BACC Balanced accuracy p. 98

BAM Bayesian Appearance Model p. 91

BIDS Brain Imaging Data Structure p. 166

CDSS Clinical Decision Support Systems pp. 28, 30, 41, 60, 61, 63, 89, 111, 113, 133, 135, 149, 154, 168, 170, 174, 175

CNN Convolutional Neural Network pp. 39, 41, 50, 54, 55, 61, 138, 139, 142, 143, 145, 147, 151, 152, 156, 158, 159, 162, 163, 169, 174

CPU Central Processing Unit p. 168

CT Computerized Tomography pp. 24, 27, 28, 177

CV Cross-Validation pp. 34, 39, 98, 99, 105, 122–124, 141, 145, 179, 180, 182, 185

DBS Deep Brain Stimulation pp. 19–21, 23, 28, 30, 37, 43, 58, 60, 61, 63, 87, 89, 91, 96, 97, 101, 102, 105, 111, 113, 115–118, 130–133, 149–152, 154–156, 160, 161, 163, 165, 170, 172, 174, 175, 177, 180, 182

DICOM Digital Imaging and Communications in Medicine p. 166

DL Deep Learning pp. 19, 32, 34, 38–41, 61, 66, 152, 154, 156, 165–167, 174, 175, 178

EEG Electroencephalography pp. 47, 52, 53

EL Ensemble Learning pp. 50, 54, 95

ET Essential Tremor pp. 46, 49

GDPR General Data Protection Regulation p. 165

GMM Gaussian Mixture Model p. 50

GPi Globus Pallidus internus pp. 20, 24, 115, 118, 119, 132

GPR Gaussian Process Regression p. 50

GPU Graphics Processing Units pp. 166, 168

HC Healthy Control pp. 93, 96, 103, 105, 182

HCTA Health Care Technology Assessment p. 168

HIPAA Health Insurance Portability and Accountability Act p. 165

HPO Hyper-parameters Optimization pp. 7, 94, 105, 121, 122, 124, 144, 145, 157, 158

HSD Honestly Significant Difference pp. 101–103

ILSVRC ImageNet Large Scale Visual Recognition Challenge pp. 39, 158, 178

IPCA Iterative PCA p. 68

kgCO2e Kilograms of carbon dioxide equivalent p. 167

kNN k-Nearest Neighbors p. 50

LDA Linear Discriminant Analysis p. 50

LDDMM Large Deformation Diffeomorphic Metric Mapping pp. 91, 95

LFP Local Field Potential pp. 47, 52, 53, 57, 172

LOOCV Leave-one-out CV p. 122

LSTM Long Short-Term Memory pp. 50, 54, 153

mse mean absolute error p. 52

MCC Matthews Correlation Coefficient p. 52

MCI Mild Control Impairment p. 92

MER Micro-electrode Recordings pp. 26, 45, 47, 53–56, 61, 132, 135, 137–140, 142, 141, 142, 145–147, 152, 154, 156, 158, 161, 174, 180

ML Machine Learning pp. 19, 23, 28, 32, 34, 36, 38–41, 43, 45, 47, 49, 52, 54, 55, 57, 58, 60, 61, 150, 151, 159–161, 166–170, 174, 175, 180

MLP Multi-Layer Perceptron pp. 50, 159

MRI Magnetic Resonance Imaging pp. 21, 23, 24, 28, 61, 89, 91, 94, 95, 97, 98, 110, 116, 119, 132, 137, 162

mse mean squared error p. 52

MSE Mean Squared Error p. 122

NB Naive Bayes p. 50

NLP Natural Language Processing pp. 39, 167

PCA Principal Component Analysis pp. 68, 94, 95, 99, 100, 120, 121, 157, 179

PD Parkinson's Disease pp. 19, 20, 46, 49, 60, 61, 67, 91–93, 96, 98, 104, 105, 108–111, 131, 138, 140, 150, 151, 154, 155, 160, 162, 182

PPCA Pairwise correlation PCA p. 68

PPMI Parkinson's Progression Markers Initiative pp. 96, 98, 101, 110

PTSE Pyramidal tract side effect pp. 23, 24, 37, 55

RBM Restricted Boltzmann Machine p. 41

RF Random Forests pp. 50, 54, 95, 105, 157

RL Reinforcement Learning p. 38

RNN Recurrent Neural Networks pp. 41, 54

ROC Receiver Operating Characteristic pp. 52, 184

RoI Region of Interest p. 95

SBM Surface-based Morphometry p. 91

Se/Sp Sensitivity and Specificity p. 52

SIFT Scale-Invariant Feature Transform p. 39

STFT Short Term Fourier Transform pp. 141, 144, 147

STN Subthalamic Nucleus pp. 20, 24, 26, 27, 30, 37, 55, 56, 61, 115, 116, 118, 119, 130, 132, 138, 139, 142, 145, 147, 152, 154, 156, 161, 174, 177, 180, 183

SVM Support Vector Machine pp. 50, 54, 95, 103, 117, 120, 151, 157, 179

UI User Interface p. 28

UTAUT Unified Theory of Acceptance of Technology p. 170

VBM Voxel Based Morphometry p. 91

VIM Ventral Intermediate nucleus of the thalamus pp. 20, 115, 118, 119

VTA Volume of Tissue Activated pp. 24, 28, 37, 55, 116, 132, 177

XGB Extreme Gradient Boosting pp. 50, 54

BIBLIOGRAPHY

[1] T. D'Albis, C. Haegelen, C. Essert, S. Fernández-Vidal, F. Lalys, and P. Jannin, « PyDBS: an automated image processing workflow for deep brain stimulation surgery », *International journal of computer assisted radiology and surgery*, vol. 10, *2*, pp. 117–128, 2015.

[2] F. Lalys, C. Haegelen, M. Mehri, S. Drapier, M. Verin, and P. Jannin, « Anatomo-clinical atlases correlate clinical data and electrode contact coordinates: application to subthalamic deep brain stimulation », *Journal of neuroscience methods*, vol. 212, *2*, pp. 297–307, 2013.

[3] C. Haegelen, C. Baumgarten, J.-F. Houvenaghel, Y. Zhao, J. Peron, S. Drapier, P. Jannin, and X. Morandi, « Functional atlases for analysis of motor and neuropsychological outcomes after medial globus pallidus and subthalamic stimulation », *PloS one*, vol. 13, *7*, e0200262, 2018.

[4] C. Baumgarten, Y. Zhao, P. Sauleau, C. Malrain, P. Jannin, and C. Haegelen, « Image-guided preoperative prediction of pyramidal tract side effect in deep brain stimulation: proof of concept and application to the pyramidal tract side effect induced by pallidal stimulation », *Journal of Medical Imaging*, vol. 3, *2*, p. 025 001, 2016.

[5] ——, « Improvement of pyramidal tract side effect prediction using a data-driven method in subthalamic stimulation », *IEEE Transactions on Biomedical Engineering*, vol. 64, *9*, pp. 2134–2141, 2016.

[6] C. Baumgarten, C. Haegelen, Y. Zhao, P. Sauleau, and P. Jannin, « Data-Driven Prediction of the Therapeutic Window during Subthalamic Deep Brain Stimulation Surgery », *Stereotactic and functional neurosurgery*, vol. 96, *3*, pp. 142–150, 2018.

[7] G. E. Alexander, « Biology of Parkinson's disease: pathogenesis and pathophysiology of a multisystem neurodegenerative disorder », *Dialogues in clinical neuroscience*, vol. 6, *3*, p. 259, 2004.

[8] R. F. Pfeiffer, « Non-motor symptoms in Parkinson's disease », *Parkinsonism & related disorders*, vol. 22, S119–S122, 2016.

[9] N. Amoroso, M. La Rocca, A. Monaco, R. Bellotti, and S. Tangaro, « Complex networks reveal early MRI markers of Parkinson's disease », *Medical image analysis*, vol. 48, pp. 12–24, 2018.

[10] A.-L. Benabid, P. Pollak, A. Louveau, S. Henry, and J. De Rougemont, « Combined (thalamotomy and stimulation) stereotactic surgery of the VIM thalamic nucleus for bilateral Parkinson disease », *Stereotactic and functional neurosurgery*, vol. 50, *1-6*, pp. 344–346, 1987.

[11] W. Hamel, J. A. Köppen, F. Alesch, A. Antonini, J. A. Barcia, H. Bergman, S. Chabardes, M. F. Contarino, P. Cornu, W. Demmel, *et al.*, « Targeting of the subthalamic nucleus for deep brain stimulation: a survey among Parkinson disease specialists », *World neurosurgery*, vol. 99, pp. 41–46, 2017.

[12] D. Drapier, S. Drapier, P. Sauleau, C. Haegelen, S. Raoul, I. Biseul, J. Peron, F. Lallement, I. Rivier, J. Reymann, *et al.*, « Does subthalamic nucleus stimulation induce apathy in Parkinson's disease? », *Journal of neurology*, vol. 253, *8*, p. 1083, 2006.

[13] K. Witt, C. Daniels, J. Reiff, P. Krack, J. Volkmann, M. O. Pinsker, M. Krause, V. Tronnier, M. Kloss, A. Schnitzler, *et al.*, « Neuropsychological and psychiatric changes after deep brain stimulation for Parkinson's disease: a randomised, multicentre study », *The Lancet Neurology*, vol. 7, *7*, pp. 605–614, 2008.

[14] X. Wang, C. Chang, N. Geng, N. Li, J. Wang, J. Ma, W. Xue, W. Zhao, H. Wu, P. Wang, *et al.*, « Long-term effects of bilateral deep brain stimulation of the subthalamic nucleus on depression in patients with Parkinson's disease », *Parkinsonism & related disorders*, vol. 15, *8*, pp. 587–591, 2009.

[15] J. Péron, I. Biseul, E. Leray, S. Vicente, F. Le Jeune, S. Drapier, D. Drapier, P. Sauleau, C. Haegelen, and M. Vérin, « Subthalamic nucleus stimulation affects fear and sadness recognition in Parkinson's disease. », *Neuropsychology*, vol. 24, *1*, p. 1, 2010.

[16] G. Tommasi, P. Krack, V. Fraix, J.-F. Le Bas, S. Chabardes, A. L. Benabid, and P. Pollak, « Pyramidal tract side effects induced by deep brain stimulation of the subthalamic nucleus », *Journal of Neurology, Neurosurgery & Psychiatry*, vol. 79, *7*, pp. 813–819, 2008.

[17] D. N. Anderson, B. Osting, J. Vorwerk, A. D. Dorval, and C. R. Butson, « Optimized programming algorithm for cylindrical and directional deep brain stimulation electrodes », *Journal of neural engineering*, vol. 15, *2*, p. 026 005, 2018.

[18] R. T. Sutton, D. Pincock, D. C. Baumgart, D. C. Sadowski, R. N. Fedorak, and K. I. Kroeker, « An overview of clinical decision support systems: benefits, risks, and strategies for success », *NPJ Digital Medicine*, vol. 3, *1*, pp. 1–10, 2020.

[19] I. Sim, P. Gorman, R. A. Greenes, R. B. Haynes, B. Kaplan, H. Lehmann, and P. C. Tang, « Clinical decision support systems for the practice of evidence-based medicine », *Journal of the American Medical Informatics Association*, vol. 8, *6*, pp. 527–534, 2001.

[20] F. S. Castro, C. Pollo, R. Meuli, P. Maeder, O. Cuisenaire, M. B. Cuadra, J.-G. Villemure, and J.-P. Thiran, « A cross validation study of deep brain stimulation targeting: from experts to atlas-based, segmentation-based and automatic registration algorithms », *IEEE transactions on medical imaging*, vol. 25, *11*, pp. 1440–1450, 2006.

[21] M. M. Chakravarty, G. Bertrand, C. P. Hodge, A. F. Sadikot, and D. L. Collins, « The creation of a brain atlas for image guided neurosurgery using serial histological data », *Neuroimage*, vol. 30, *2*, pp. 359–376, 2006.

[22] S. Ourselin, E. Bardinet, D. Dormont, G. Malandain, A. Roche, N. Ayache, D. Tandé, K. Parain, and J. Yelnik, « Fusion of histological sections and MR images: towards the construction of an atlas of the human basal ganglia », *in International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2001, pp. 743–751.

[23] E. Bardinet, D. Dormont, G. Malandain, M. Bhattacharjee, B. Pidoux, C. Saleh, P. Cornu, N. Ayache, Y. Agid, and J. Yelnik, « Retrospective cross-evaluation of an histological and deformable 3D atlas of the basal ganglia on series of Parkinsonian patients treated by deep brain stimulation », *in International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2005, pp. 385–393.

[24] T. Guo, K. W. Finnis, A. G. Parrent, and T. M. Peters, « Development and application of functional databases for planning deep-brain neurosurgical procedures », *in International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2005, pp. 835–842.

[25] P.-F. D'Haese, E. Cetinkaya, P. E. Konrad, C. Kao, and B. M. Dawant, « Computer-aided placement of deep brain stimulators: from planningto intraoperative guidance », *IEEE transactions on medical imaging*, vol. 24, *11*, pp. 1469–1478, 2005.

[26] A. L. Samuel, « Some studies in machine learning using the game of checkers », *IBM Journal of research and development*, vol. 3, *3*, pp. 210–229, 1959.

[27] E. Celtikci, « A systematic review on machine learning in neurosurgery: the future of decision-making in patient care », *Turk Neurosurg*, vol. 28, *2*, pp. 167–173, 2018.

[28] O. Vinyals, I. Babuschkin, J. Chung, M. Mathieu, M. Jaderberg, W. Czarnecki, A. Dudzik, A. Huang, P. Georgiev, R. Powell, T. Ewalds, D. Horgan, M. Kroiss, I. Danihelka, J. Agapiou, J. Oh, V. Dalibard, D. Choi, L. Sifre, Y. Sulsky, S. Vezhnevets, J. Molloy, T. Cai, D. Budden, T. Paine, C. Gulcehre, Z. Wang, T. Pfaff, T. Pohlen, D. Yogatama, J. Cohen, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, C. Apps, K. Kavukcuoglu, D. Hassabis, and D. Silver, *AlphaStar: Mastering the Real-Time Strategy Game StarCraft II*, `https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/`, 2019.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, « ImageNet Large Scale Visual Recognition Challenge », *International Journal of Computer Vision (IJCV)*, vol. 115, *3*, pp. 211–252, 2015. DOI: `10.1007/s11263-015-0816-y`.

[30] K. Simonyan and A. Zisserman, « Very deep convolutional networks for large-scale image recognition », *arXiv preprint arXiv:1409.1556*, 2014.

[31] A. Graves, A.-r. Mohamed, and G. Hinton, « Speech recognition with deep recurrent neural networks », *in 2013 IEEE international conference on acoustics, speech and signal processing*, IEEE, 2013, pp. 6645–6649.

[32] A. Radford, L. Metz, and S. Chintala, « Unsupervised representation learning with deep convolutional generative adversarial networks », *arXiv preprint arXiv:1511.06434*, 2015.

[33] S. Purushotham, C. Meng, Z. Che, and Y. Liu, « Benchmarking deep learning models on large healthcare datasets », *Journal of biomedical informatics*, vol. 83, pp. 112–134, 2018.

[34] R. Perrault, Y. Shoham, E. Brynjolfsson, J. Clark, J. Etchemendy, B. Grosz, T. Lyons, J. Manyika, S. Mishra, and J. C. Niebles, « The AI Index 2019 Annual Report », *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA*, 2019.

[35] C. Villani, Y. Bonnet, B. Rondepierre, *et al.*, *For a meaningful artificial intelligence: Towards a French and European strategy*. Conseil national du numérique, 2018.

[36] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, « Dermatologist-level classification of skin cancer with deep neural networks », *nature*, vol. 542, *7639*, pp. 115–118, 2017.

[37] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, « Deep learning for healthcare: review, opportunities and challenges », *Briefings in bioinformatics*, vol. 19, *6*, pp. 1236–1246, 2018.

[38] J. T. Senders, M. M. Zaki, A. V. Karhade, B. Chang, W. B. Gormley, M. L. Broekman, T. R. Smith, and O. Arnaout, « An introduction and overview of machine learning in neurosurgical care », *Acta neurochirurgica*, vol. 160, *1*, pp. 29–38, 2018.

[39] Q. D. Buchlak, N. Esmaili, J.-C. Leveque, F. Farrokhi, C. Bennett, M. Piccardi, and R. K. Sethi, « Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review », *Neurosurgical review*, pp. 1–19, 2019.

[40] J. T. Senders, O. Arnaout, A. V. Karhade, H. H. Dasenbrock, W. B. Gormley, M. L. Broekman, and T. R. Smith, « Natural and artificial intelligence in neurosurgery: a systematic review », *Neurosurgery*, vol. 83, *2*, pp. 181–192, 2018.

[41] A. Orozco, M. Alvarez, E. Guijarro, and G. Castellanos, « Identification of spike sources using proximity analysis through hidden Markov models », *in 2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2006, pp. 5555–5558.

[42] A. Muniz, W. Liu, H. Liu, K. Lyons, R. Pahwa, F. Nobre, and J. Nadal, « Assessment of the effects of subthalamic stimulation in Parkinson disease patients by artificial neural network », *in 2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2009, pp. 5673–5676.

[43] S. Wong, G. Baltuch, J. Jaggi, and S. Danish, « Functional localization and visualization of the subthalamic nucleus from microelectrode recordings acquired during DBS surgery with unsupervised machine learning », *Journal of neural engineering*, vol. 6, *2*, p. 026 006, 2009.

[44] D. Wu, K. Warwick, Z. Ma, M. N. Gasson, J. G. Burgess, S. Pan, and T. Z. Aziz, « Prediction of Parkinson's disease tremor onset using a radial basis function neural network based on particle swarm optimization », *International journal of neural systems*, vol. 20, *02*, pp. 109–116, 2010.

[45] P. Guillén, F. Martinez-de-Pison, R. Sanchez, M. Argáez, and L. Velázquez, « Characterization of subcortical structures during deep brain stimulation utilizing support vector machines », *in 2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2011, pp. 7949–7952.

[46] P. Shukla, I. Basu, D. Graupe, D. Tuninetti, and K. V. Slavin, « A neural network-based design of an on-off adaptive control for Deep Brain Stimulation in movement disorders », *in 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2012, pp. 4140–4143.

[47] C. Loukas and P. Brown, « A PC-based system for predicting movement from deep brain signals in Parkinson's disease », *Computer methods and programs in biomedicine*, vol. 107, *1*, pp. 36–44, 2012.

[48] H. Jiang, J. J. Zhang, A. Hebb, and M. H. Mahoor, « Time-frequency analysis of brain electrical Signals for behvior recognition in patients with Parkinson's disease », *in 2013 Asilomar Conference on Signals, Systems and Computers*, IEEE, 2013, pp. 1843–1847.

[49] S. Niketeghad, A. O. Hebb, J. Nedrud, S. J. Hanrahan, and M. H. Mahoor, « Single trial behavioral task classification using subthalamic nucleus local field potential signals », *in 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2014, pp. 3793–3796.

[50] A. T. Connolly, W. F. Kaemmerer, S. Dani, S. R. Stanslaski, E. Panken, M. D. Johnson, and T. Denison, « Guiding deep brain stimulation contact selection using local field potentials sensed by a chronically implanted device in Parkinson's disease patients », *in 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER)*, IEEE, 2015, pp. 840–843.

[51] R. R. Shamir, T. Dolber, A. M. Noecker, B. L. Walter, and C. C. McIntyre, « Machine learning approach to optimizing combined stimulation and medication therapies for Parkinson's disease », *Brain stimulation*, vol. 8, *6*, pp. 1025–1032, 2015.

[52] V. Rajpurohit, S. F. Danish, E. L. Hargreaves, and S. Wong, « Optimizing computational feature sets for subthalamic nucleus localization in DBS surgery with feature selection », *Clinical Neurophysiology*, vol. 126, *5*, pp. 975–982, 2015.

[53] J. Kim, Y. Duchin, H. Kim, J. Vitek, N. Harel, and G. Sapiro, « Robust prediction of clinical deep brain stimulation target structures via the estimation of influential high-field MR atlases », *in International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2015, pp. 587–594.

[54] N. Khobragade, D. Graupe, and D. Tuninetti, « Towards fully automated closed-loop Deep Brain Stimulation in Parkinson's disease patients: a LAMSTAR-based tremor predictor », *in 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2015, pp. 2616–2619.

[55] S. A. Yohanandan, M. Jones, R. Peppard, J. L. Tan, H. J. McDermott, and T. Perera, « Evaluating machine learning algorithms estimating tremor severity ratings on the Bain–Findley scale », *Measurement Science and Technology*, vol. 27, *12*, p. 125 702, 2016.

[56] Y. Liu and B. M. Dawant, « Multi-modal learning-based pre-operative targeting in deep brain stimulation procedures », *in 2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)*, IEEE, 2016, pp. 17–20.

[57] P. Angeles, Y. Tai, N. Pavese, S. Wilson, and R. Vaidyanathan, « Automated assessment of symptom severity changes during deep brain stimulation (DBS) therapy for Parkinson's disease », *in 2017 International Conference on Rehabilitation Robotics (ICORR)*, IEEE, 2017, pp. 1512–1517.

[58] K. Kostoglou, K. P. Michmizos, P. Stathis, D. Sakas, K. S. Nikita, and G. D. Mitsis, « Classification and prediction of clinical improvement in deep brain stimulation from intraoperative microelectrode recordings », *IEEE Transactions on Biomedical Engineering*, vol. 64, *5*, pp. 1123–1130, 2016.

[59] B. C. Houston, M. C. Thompson, J. G. Ojemann, A. L. Ko, and H. J. Chizeck, « Classifier-based closed-loop deep brain stimulation for essential tremor », *in 2017 8th International IEEE/EMBS Conference on Neural Engineering (NER)*, IEEE, 2017, pp. 316–320.

[60] P. Guillén-Rondon and M. D. Robinson, « Deep brain stimulation signal classification using deep belief networks », *in 2016 International Conference on Computational Science and Computational Intelligence (CSCI)*, IEEE, 2016, pp. 155–158.

[61] F. Milletari, S.-A. Ahmadi, C. Kroll, A. Plate, V. Rozanski, J. Maiostre, J. Levin, O. Dietrich, B. Ertl-Wagner, K. Bötzel, *et al.*, « Hough-CNN: deep learning for segmentation of deep brain regions in MRI and ultrasound », *Computer Vision and Image Understanding*, vol. 164, pp. 92–102, 2017.

[62] D. Valsky, O. Marmor-Levin, M. Deffains, R. Eitan, K. T. Blackwell, H. Bergman, and Z. Israel, « Stop! border ahead: A utomatic detection of subthalamic exit during deep brain stimulation surgery », *Movement Disorders*, vol. 32, *1*, pp. 70–79, 2017.

[63] A. Mohammed, M. Zamani, R. Bayford, and A. Demosthenous, « Toward on-demand deep brain stimulation using online Parkinson's disease prediction driven by dynamic detection », *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 25, *12*, pp. 2441–2452, 2017.

[64] H. M. Golshan, A. O. Hebb, S. J. Hanrahan, J. Nedrud, and M. H. Mahoor, « A hierarchical structure for human behavior classification using STN local field potentials », *Journal of neuroscience methods*, vol. 293, pp. 254–263, 2018.

[65] M. Khosravi, S. F. Atashzar, G. Gilmore, M. S. Jog, and R. V. Patel, « Electrophysiological signal processing for intraoperative localization of subthalamic nucleus during deep brain stimulation surgery », *in 2018 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, IEEE, 2018, pp. 424–428.

[66] R. LeMoyne, T. Mastroianni, C. McCandless, C. Currivan, D. Whiting, and N. Tomycz, « Implementation of a smartphone as a wearable and wireless accelerometer and gyroscope platform for ascertaining deep brain stimulation treatment efficacy of Parkinson's disease through machine learning classification », *Advances in Parkinson's Disease*, vol. 7, *2*, pp. 19–30, 2018.

[67] H. D. V. Cardona, M. A. Álvarez, and Á. A. Orozco, « Multi-task learning for subthalamic nucleus identification in deep brain stimulation », *International Journal of Machine Learning and Cybernetics*, vol. 9, *7*, pp. 1181–1192, 2018.

[68] N. Khobragade, D. Tuninetti, and D. Graupe, « On the need for adaptive learning in on-demand Deep Brain Stimulation for Movement Disorders », *in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018, pp. 2190–2193.

[69] F. H. M. Oliveira, A. R. Machado, and 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, 2018.

[70] S. A. Shah, G. Tinkhauser, C. C. Chen, S. Little, and P. Brown, « Parkinsonian tremor detection from subthalamic nucleus local field potentials for closed-loop deep brain stimulation », *in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018, pp. 2320–2324.

[71] L. Yao, P. Brown, and M. Shoaran, « Resting Tremor Detection in Parkinson's Disease with Machine Learning and Kalman Filtering », *in 2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, IEEE, 2018, pp. 1–4.

[72] H. M. Golshan, A. O. Hebb, J. Nedrud, and M. H. Mahoor, « Studying the Effects of Deep Brain Stimulation and Medication on the Dynamics of STN-LFP Signals for Human Behavior Analysis », *in 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2018, pp. 4720–4723.

[73] T. Wang, M. Shoaran, and A. Emami, « Towards Adaptive Deep Brain Stimulation in Parkinson'S Disease: Lfp-Based Feature Analysis and Classification », *in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2018, pp. 2536–2540.

[74] B. Houston, M. Thompson, A. Ko, and H. Chizeck, « A machine-learning approach to volitional control of a closed-loop deep brain stimulation system », *Journal of neural engineering*, vol. 16, *1*, p. 016 004, 2018.

[75] M. Koch, V. Geraedts, H. Wang, M. Tannemaat, and T. Bäck, « Automated Machine Learning for EEG-Based Classification of Parkinson's Disease Patients », *in 2019 IEEE International Conference on Big Data (Big Data)*, IEEE, 2019, pp. 4845–4852.

[76] J. Kim, Y. Duchin, R. R. Shamir, R. Patriat, J. Vitek, N. Harel, and G. Sapiro, « Automatic localization of the subthalamic nucleus on patient-specific clinical MRI by incorporating 7 T MRI and machine learning: Application in deep brain stimulation », *Human brain mapping*, vol. 40, *2*, pp. 679–698, 2019.

[77] Y. Chen, C. Gong, H. Hao, Y. Guo, S. Xu, Y. Zhang, G. Yin, X. Cao, A. Yang, F. Meng, *et al.*, « Automatic Sleep Stage Classification Based on Subthalamic Local Field Potentials », *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, *2*, pp. 118–128, 2019.

[78] H. Tan, J. Debarros, S. He, A. Pogosyan, T. Z. Aziz, Y. Huang, S. Wang, L. Timmermann, V. Visser-Vandewalle, D. J. Pedrosa, *et al.*, « Decoding voluntary movements and postural tremor based on thalamic LFPs as a basis for closed-loop stimulation for essential tremor », *Brain stimulation*, vol. 12, *4*, pp. 858–867, 2019.

[79] S.-C. Park, J. H. Cha, S. Lee, W. Jang, C. S. Lee, and J. K. Lee, « Deep Learning-Based Deep Brain Stimulation Targeting and Clinical Applications. », *Frontiers in neuroscience*, vol. 13, pp. 1128–1128, 2019.

[80] R. LeMoyne, T. Mastroianni, C. McCandless, D. Whiting, and N. Tomycz, « Evaluation of Machine Learning Algorithms for Classifying Deep Brain Stimulation Respective of 'On'and 'Off'Status », *in 2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)*, IEEE, 2019, pp. 483–488.

[81] O. Klempıř, R. Krupička, E. Bakštein, and R. Jech, « Identification of Microrecording Artifacts with Wavelet Analysis and Convolutional Neural Network: An Image Recognition Approach », *Measurement Science Review*, vol. 19, *5*, pp. 222–231, 2019.

[82] M. Stuart, C. S. Wickramasinghe, D. L. Marino, D. Kumbhare, K. Holloway, and M. Manic, « Machine Learning for Deep Brain Stimulation Efficacy using Dense Array EEG », *in 2019 12th International Conference on Human System Interaction (HSI)*, IEEE, 2019, pp. 143–150.

[83] J. G. Habets, A. A. Duits, L. C. Sijben, B. De Greef, A. Mulders, Y. Temel, M. L. Kuijf, P. L. Kubben, C. Herff, and M. L. Janssen, « Machine learning prediction of motor response after deep brain stimulation in Parkinson's disease », *medRxiv*, p. 19 006 841, 2019.

[84] C. Camara, N. P. Subramaniyam, K. Warwick, L. Parkkonen, T. Aziz, and E. Pereda, « Non-Linear Dynamical Analysis of Resting Tremor for Demand-Driven Deep Brain Stimulation », *Sensors*, vol. 19, *11*, p. 2507, 2019.

[85] A. Singer, C. Zhang, T. Wang, S. Qiu, D. Li, Y. Du, Z.-P. Liang, P. Herman, B. Sun, and Y. Feng, « Post-operative electrode placement prediction in deep brain stimulation using support vector regression », *in Proceedings of the Third International Symposium on Image Computing and Digital Medicine*, 2019, pp. 202–207.

[86] C. Bermudez, W. Rodriguez, Y. Huo, A. E. Hainline, R. Li, R. Shults, P. D. D'Haese, P. E. Konrad, B. M. Dawant, and B. A. Landman, « Towards machine learning prediction of deep brain stimulation (DBS) intra-operative efficacy maps », *in Medical Imaging 2019: Image Processing*, International Society for Optics and Photonics, vol. 10949, 2019, p. 1 094 922.

[87] K. A. Ciecierski and T. Mandat, « Unsupervised machine learning in classification of neurobiological data », *in Intelligent Methods and Big Data in Industrial Applications*, Springer, 2019, pp. 203–212.

[88] A. Mohammed, R. Bayford, and A. Demosthenous, « A framework for adapting deep brain stimulation using Parkinsonian state estimates », *Frontiers in Neuroscience*, vol. 14, p. 499, 2020.

[89] M. Hosny, M. Zhu, W. Gao, and Y. Fu, « A novel deep LSTM network for artifacts detection in microelectrode recordings », *Biocybernetics and Biomedical Engineering*, 2020.

[90] F. Farrokhi, Q. D. Buchlak, M. Sikora, N. Esmaili, M. Marsans, P. McLeod, J. Mark, E. Cox, C. Bennett, and J. Carlson, « Investigating risk factors and predicting complications in deep brain stimulation surgery with machine learning algorithms », *World Neurosurgery*, vol. 134, e325–e338, 2020.

[91] D. Valsky, K. T. Blackwell, I. Tamir, R. Eitan, H. Bergman, and Z. Israel, « Real-time machine learning classification of pallidal borders during deep brain stimulation surgery », *Journal of Neural Engineering*, vol. 17, *1*, p. 016 021, 2020.

[92] J. S. Baxter, E. Maguet, and P. Jannin, « Localisation of the subthalamic nucleus in MRI via convolutional neural networks for deep brain stimulation planning », *in Medical Imaging 2020: Image-Guided Procedures, Robotic Interventions, and Modeling*, International Society for Optics and Photonics, vol. 11315, 2020, p. 113150M.

[93] T. A. Mestre, S. Eberly, C. Tanner, D. Grimes, A. E. Lang, D. Oakes, and C. Marras, « Reproducibility of data-driven Parkinson's disease subtypes for clinical research », *Parkinsonism & Related Disorders*, vol. 56, pp. 102–106, 2018.

[94] B. K. Beaulieu-Jones, D. R. Lavage, J. W. Snyder, J. H. Moore, S. A. Pendergrass, and C. R. Bauer, « Characterizing and managing missing structured data in electronic health records: data analysis », *JMIR medical informatics*, vol. 6, *1*, e11, 2018.

[95] S. Srivastava, S. Soman, A. Rai, and P. K. Srivastava, « Deep learning for health informatics: Recent trends and future directions », *in 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, IEEE, 2017, pp. 1665–1670.

[96] B. Efron, « Missing data, imputation, and the bootstrap », *Journal of the American Statistical Association*, vol. 89, *426*, pp. 463–475, 1994.

[97] S. Das, S. Datta, and B. B. Chaudhuri, « Handling data irregularities in classification: Foundations, trends, and future challenges », *Pattern Recognition*, vol. 81, pp. 674–693, 2018.

[98] A. Sánchez-Morales, J.-L. Sancho-Gómez, J.-A. Martınez-Garcıa, and A. R. Figueiras-Vidal, « Improving deep learning performance with missing values via deletion and compensation », *Neural Computing and Applications*, pp. 1–12, 2019.

[99] U. Hwang, S. Choi, H.-B. Lee, and S. Yoon, « Adversarial training for disease prediction from electronic health records with missing data », *arXiv preprint arXiv:1711.04126*, 2017.

[100] C. Xiao, E. Choi, and J. Sun, « Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review », *Journal of the American Medical Informatics Association*, vol. 25, *10*, pp. 1419–1428, 2018.

[101] J. M. Kishton and K. F. Widaman, « Unidimensional versus domain representative parceling of questionnaire items: An empirical example », *Educational and Psychological Measurement*, vol. 54, *3*, pp. 757–765, 1994.

[102] P. Hagell and M. H. Nilsson, « The 39-item Parkinson's Disease Questionnaire (PDQ-39): is it a unidimensional construct? », *Therapeutic Advances in Neurological Disorders*, vol. 2, *4*, pp. 205–214, 2009.

[103] V. Abedi, M. K. Shivakumar, P. Lu, R. Hontecillas, A. Leber, M. Ahuja, A. E. Ulloa, J. M. Shellenberger, and J. Bassaganya-Riera, « Latent-Based Imputation of Laboratory Measures from Electronic Health Records: Case for Complex Diseases », *bioRxiv*, p. 275 743, 2018.

[104] B. K. Beaulieu-Jones and J. H. Moore, « Missing data imputation in the electronic health record using deeply learned autoencoders », *in Pacific Symposium on Biocomputing 2017*, World Scientific, 2017, pp. 207–218.

[105] G. E. Hinton and R. R. Salakhutdinov, « Reducing the dimensionality of data with neural networks », *science*, vol. 313, *5786*, pp. 504–507, 2006.

[106] M. Abdella and T. Marwala, « The use of genetic algorithms and neural networks to approximate missing data in database », *in Computational Cybernetics, 2005. ICCC 2005. IEEE 3rd International Conference on*, IEEE, 2005, pp. 207–212.

[107] R. Miotto, L. Li, B. A. Kidd, and J. T. Dudley, « Deep patient: an unsupervised representation to predict the future of patients from the electronic health records », *Scientific reports*, vol. 6, p. 26 094, 2016.

[108] P. D. Allison, *Missing data*. 1999.

[109] S. Dray and J. Josse, « Principal component analysis with missing values: a comparative survey of methods », *Plant Ecology*, vol. 216, *5*, pp. 657–667, 2015.

[110] H. A. Kiers, « Weighted least squares fitting using ordinary least squares algorithms », *Psychometrika*, vol. 62, *2*, pp. 251–266, 1997.

[111] J. Josse, F. Husson, and J. Pagès, « Gestion des données manquantes en analyse en composantes principales », *Journal de la Société Française de Statistique*, vol. 150, *2*, pp. 28–51, 2009.

[112] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, « Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion », *Journal of Machine Learning Research*, vol. 11, *Dec*, pp. 3371–3408, 2010.

[113] S. Chandar, M. M. Khapra, H. Larochelle, and B. Ravindran, « Correlational neural networks », *Neural computation*, vol. 28, *2*, pp. 257–285, 2016.

[114] H. Zhang, P. Xie, and E. Xing, « Missing value imputation based on deep generative models », *arXiv preprint arXiv:1808.01684*, 2018.

[115] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, « Densely connected convolutional networks », *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[116] A. F. Costa, M. S. Santos, J. P. Soares, and P. H. Abreu, « Missing data imputation via denoising autoencoders: the untold story », *in International Symposium on Intelligent Data Analysis*, Springer, 2018, pp. 87–98.

[117] N. Jaques, S. Taylor, A. Sano, and R. Picard, « Multimodal Autoencoder: A Deep Learning Approach to Filling In Missing Sensor Data and Enabling Better Mood Prediction », *in Proc. International Conference on Affective Computing and Intelligent Interaction (ACII), San Antonio, Texas*, 2017.

[118] K. Marek, D. Jennings, S. Lasch, A. Siderowf, C. Tanner, T. Simuni, C. Coffey, K. Kieburtz, E. Flagg, S. Chowdhury, *et al.*, « The parkinson progression marker initiative (PPMI) », *Progress in neurobiology*, vol. 95, *4*, pp. 629–635, 2011.

[119] C. G. Goetz, S. Luo, L. Wang, B. C. Tilley, N. R. LaPelle, and G. T. Stebbins, « Handling missing values in the MDS-UPDRS », *Movement Disorders*, vol. 30, *12*, pp. 1632–1638, 2015.

[120] J. de La Torre, D. Puig, and A. Valls, « Weighted kappa loss function for multi-class classification of ordinal data in deep learning », *Pattern Recognition Letters*, vol. 105, pp. 144–154, 2018.

[121] P. Mahlknecht, K. Seppi, and W. Poewe, « The concept of prodromal Parkinson's disease », *Journal of Parkinson's disease*, vol. 5, *4*, pp. 681–697, 2015.

[122] M. A. Thenganatt and J. Jankovic, « Parkinson disease subtypes », *JAMA neurology*, vol. 71, *4*, pp. 499–504, 2014.

[123] B. S. Connolly and A. E. Lang, « Pharmacological treatment of Parkinson disease: a review », *Jama*, vol. 311, *16*, pp. 1670–1683, 2014.

[124] P. Limousin and I. Martinez-Torres, « Deep brain stimulation for Parkinson's disease », *Neurotherapeutics*, vol. 5, *2*, pp. 309–319, 2008.

[125] P. Limousin and T. Foltynie, « Long-term outcomes of deep brain stimulation in Parkinson disease », *Nature Reviews Neurology*, vol. 15, *4*, pp. 234–242, 2019.

[126] M. Delenclos, D. R. Jones, P. J. McLean, and R. J. Uitti, « Biomarkers in Parkinson's disease: advances and strategies », *Parkinsonism & related disorders*,

[127] S. J. Kish, K. Shannak, and O. Hornykiewicz, « Uneven pattern of dopamine loss in the striatum of patients with idiopathic Parkinson's disease », *New England Journal of Medicine*, vol. 318, *14*, pp. 876–880, 1988.

[128] P. Griffiths, R. Perry, and A. Crossman, « A detailed anatomical analysis of neurotransmitter receptors in the putamen and caudate in Parkinson's disease and Alzheimer's disease », *Neuroscience letters*, vol. 169, *1-2*, pp. 68–72, 1994.

[129] P. Péran, F. Nemmi, and G. Barbagallo, « Brain Morphometry: Parkinson's Disease », *in Brain Morphometry*, Springer, 2018, pp. 267–277.

[130] P. Tuite, « Magnetic resonance imaging as a potential biomarker for Parkinson's disease », *Translational Research*, vol. 175, pp. 4–16, 2016.

[131] L. Bergouignan, M. Chupin, Y. Czechowska, S. Kinkingnéhun, C. Lemogne, G. Le Bastard, M. Lepage, L. Garnero, O. Colliot, and P. Fossati, « Can voxel based morphometry, manual segmentation and automated segmentation equally detect hippocampal volume differences in acute depression? », *Neuroimage*, vol. 45, *1*, pp. 29–37, 2009.

[132] C. Davatzikos, « Why voxel-based morphometric analysis should be used with great caution when characterizing group differences », *Neuroimage*, vol. 23, *1*, pp. 17–20, 2004.

[133] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes, « Computing large deformation metric mappings via geodesic flows of diffeomorphisms », *International journal of computer vision*, vol. 61, *2*, pp. 139–157, 2005.

[134] L. A. Berner, Z. Wang, M. Stefan, S. Lee, Z. Huo, M. Cyr, and R. Marsh, « Subcortical shape abnormalities in bulimia nervosa », *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 2019.

[135] S. J. van den Bogaard, E. M. Dumas, L. Ferrarini, J. Milles, M. A. van Buchem, J. van der Grond, and R. A. Roos, « Shape analysis of subcortical nuclei in Huntington's disease, global versus local atrophy—Results from the TRACK-HD study », *Journal of the neurological sciences*, vol. 307, *1-2*, pp. 60–68, 2011.

[136] G. Gerig, M. Styner, M. E. Shenton, and J. A. Lieberman, « Shape versus size: Improved understanding of the morphology of brain structures », *in International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2001, pp. 24–32.

[137] R. A. Menke, K. Szewczyk-Krolikowski, S. Jbabdi, M. Jenkinson, K. Talbot, C. E. Mackay, and M. Hu, « Comprehensive morphometry of subcortical grey matter structures in early-stage Parkinson's disease », *Human brain mapping*, vol. 35, *4*, pp. 1681–1690, 2014.

[138] X. Tang, D. Holland, A. M. Dale, L. Younes, M. I. Miller, and A. D. N. Initiative, « Shape abnormalities of subcortical and ventricular structures in mild cognitive impairment and Alzheimer's disease: detecting, quantifying, and predicting », *Human brain mapping*, vol. 35, *8*, pp. 3701–3725, 2014.

[139] B. S. Wade, S. H. Joshi, T. Pirnia, A. M. Leaver, R. P. Woods, P. M. Thompson, R. Espinoza, and K. L. Narr, « Random forest classification of depression status based on subcortical brain morphometry following electroconvulsive therapy », *in 2015 IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, IEEE, 2015, pp. 92–96.

[140] A. Garg, S. Appel-Cresswell, K. Popuri, M. J. McKeown, and M. F. Beg, « Morphological alterations in the caudate, putamen, pallidum, and thalamus in Parkinson's disease », *Frontiers in Neuroscience*, vol. 9, p. 101, 2015.

[141] F. Nemmi, U. Sabatini, O. Rascol, and P. Péran, « Parkinson's disease and local atrophy in subcortical nuclei: insight from shape analysis », *Neurobiology of aging*, vol. 36, *1*, pp. 424–433, 2015.

[142] C. Owens-Walton, D. Jakabek, X. Li, F. A. Wilkes, M. Walterfang, D. Velakoulis, D. Van Westen, J. C. Looi, and O. Hansson, « Striatal changes in Parkinson disease: An investigation of morphology, functional connectivity and their relationship to clinical symptoms », *Psychiatry Research: Neuroimaging*, vol. 275, pp. 5–13, 2018.

[143] E. Mak, N. Bergsland, M. Dwyer, R. Zivadinov, and N. Kandiah, « Subcortical atrophy is associated with cognitive impairment in mild Parkinson disease: a combined investigation of volumetric changes, cortical thickness, and vertex-based shape analysis », *American Journal of Neuroradiology*, vol. 35, *12*, pp. 2257–2264, 2014.

[144] H. Foo, E. Mak, T. Yong, M. Wen, R. Chander, W. Au, Y. Sitoh, L. Tan, and N. Kandiah, « Progression of subcortical atrophy in mild Parkinson's disease and its impact on cognition », *European journal of neurology*, vol. 24, *2*, pp. 341–348, 2017.

[145] A. R. Khan, N. M. Hiebert, A. Vo, B. T. Wang, A. M. Owen, K. N. Seergobin, and P. A. MacDonald, « Biomarkers of Parkinson's disease: Striatal sub-regional structural morphometry and diffusion MRI », *NeuroImage: Clinical*, vol. 21, p. 101 597, 2019.

[146] A. R. Khan, L. Wang, and M. F. Beg, « FreeSurfer-initiated fully-automated subcortical brain segmentation in MRI using large deformation diffeomorphic metric mapping », *Neuroimage*, vol. 41, *3*, pp. 735–746, 2008.

[147] A. Garg, D. Wong, K. Popuri, K. J. Poskitt, K. Fitzpatrick, B. Bjornson, R. E. Grunau, and M. F. Beg, « Manually segmented template library for 8-year-old pediatric brain MRI data with 16 subcortical structures », *Journal of Medical Imaging*, vol. 1, *3*, p. 034 502, 2014.

[148] L. Wang, A. Khan, J. G. Csernansky, B. Fischl, M. I. Miller, J. C. Morris, and M. F. Beg, « Fully-automated, multi-stage hippocampus mapping in very mild Alzheimer disease », *Hippocampus*, vol. 19, *6*, pp. 541–548, 2009.

[149] S. U. Ansari, « Validation of FS+ LDDMM by automatic segmentation of caudate nucleus in brain MRI », *in Proceedings of the 8th International Conference on Frontiers of Information Technology*, ACM, 2010, p. 10.

[150] J. G. Hentz, S. H. Mehta, H. A. Shill, E. Driver-Dunckley, T. G. Beach, and C. H. Adler, « Simplified conversion method for unified Parkinson's disease rating scale motor examinations », *Movement Disorders*, vol. 30, *14*, pp. 1967–1970, 2015.

[151] Y. Xiao, V. Fonov, M. M. Chakravarty, S. Beriault, F. Al Subaie, A. Sadikot, G. B. Pike, G. Bertrand, and D. L. Collins, « A dataset of multi-contrast population-

averaged brain MRI atlases of a Parkinson's disease cohort », *Data in brief*, vol. 12, pp. 370–379, 2017.

[152] C. S. Kubu, « The role of a neuropsychologist on a movement disorders deep brain stimulation team », *Archives of Clinical Neuropsychology*, vol. 33, *3*, pp. 365–374, 2018.

[153] G. Kleiner-Fisman, J. Herzog, D. N. Fisman, F. Tamma, K. E. Lyons, R. Pahwa, A. E. Lang, and G. Deuschl, « Subthalamic nucleus deep brain stimulation: summary and meta-analysis of outcomes », *Movement disorders: official journal of the Movement Disorder Society*, vol. 21, *S14*, S290–S304, 2006.

[154] A. I. Tröster, J. Jankovic, M. Tagliati, D. Peichel, and M. S. Okun, « Neuropsychological outcomes from constant current deep brain stimulation for Parkinson's disease », *Movement Disorders*, vol. 32, *3*, pp. 433–440, 2017.

[155] J.-H. Heo, K.-M. Lee, S. H. Paek, M.-J. Kim, J.-Y. Lee, J.-Y. Kim, S.-Y. Cho, Y. H. Lim, M.-R. Kim, S. Y. Jeong, *et al.*, « The effects of bilateral subthalamic nucleus deep brain stimulation (STN DBS) on cognition in Parkinson disease », *Journal of the neurological sciences*, vol. 273, *1-2*, pp. 19–24, 2008.

[156] L. Mugge, B. Krafcik, G. Pontasch, A. Alnemari, J. Neimat, and D. Gaudin, « A review of biomarkers use in Parkinson with deep brain stimulation: a successful past promising a bright future », *World Neurosurgery*, vol. 123, pp. 197–207, 2019.

[157] A. E. Lang, J.-L. Houeto, P. Krack, C. Kubu, K. E. Lyons, E. Moro, W. Ondo, R. Pahwa, W. Poewe, A. I. Tröster, *et al.*, « Deep brain stimulation: preoperative issues », *Movement disorders: official journal of the Movement Disorder Society*, vol. 21, *S14*, S171–S196, 2006.

[158] P. Pollak, « Deep brain stimulation for Parkinson's disease–patient selection », *in Handbook of clinical neurology*, vol. 116, Elsevier, 2013, pp. 97–105.

[159] J. T. Senders, P. C. Staples, A. V. Karhade, M. M. Zaki, W. B. Gormley, M. L. Broekman, T. R. Smith, and O. Arnaout, « Machine learning and neurosurgical outcome prediction: a systematic review », *World neurosurgery*, vol. 109, pp. 476–486, 2018.

[160] M. G. Rizzone, T. Martone, R. Balestrino, and L. Lopiano, « Genetic background and outcome of Deep Brain Stimulation in Parkinson's disease », *Parkinsonism & related disorders*, vol. 64, pp. 8–19, 2019.

[161] J. L. Jaggi, A. Umemura, H. I. Hurtig, A. D. Siderowf, A. Colcher, M. B. Stern, and G. H. Baltuch, « Bilateral stimulation of the subthalamic nucleus in Parkinson's disease: surgical efficacy and prediction of outcome », *Stereotactic and functional neurosurgery*, vol. 82, *2-3*, pp. 104–114, 2004.

[162] M. Welter, J. Houeto, S. Tezenas du Montcel, V. Mesnage, A. Bonnet, B. Pillon, I. Arnulf, B. Pidoux, D. Dormont, P. Cornu, *et al.*, « Clinical predictive factors of subthalamic stimulation in Parkinson's disease », *Brain*, vol. 125, *3*, pp. 575–583, 2002.

[163] S. Watanabe, K. Suenaga, A. Yamamoto, K. Abe, N. Kotoura, R. Ishikura, S. Hirota, and H. Yoshikawa, « Correlation of subthalamic nuclei T2 relaxation times with neuropsychological symptoms in patients with Parkinson's disease », *Journal of the neurological sciences*, vol. 315, *1-2*, pp. 96–99, 2012.

[164] T. Lönnfors-Weitzel, T. Weitzel, J. Slotboom, C. Kiefer, C. Pollo, M. Schüpbach, M. Oertel, A. Kaelin, and R. Wiest, « T2-relaxometry predicts outcome of DBS in idiopathic Parkinson's disease », *NeuroImage: Clinical*, vol. 12, pp. 832–837, 2016.

[165] T. Ballarini, K. Mueller, F. Albrecht, F. Růžička, O. Bezdicek, E. Růžička, J. Roth, J. Vymazal, R. Jech, and M. L. Schroeter, « Regional gray matter changes and age predict individual treatment response in Parkinson's disease », *NeuroImage: Clinical*, vol. 21, p. 101 636, 2019.

[166] L. J. Weinkle, B. Hoyt, J. A. Thompson, S. Sillau, J. Tanabe, J. Honce, and O. Klepitskaya, « Association of MRI Measurements with Cognitive Outcomes After STN-DBS in Parkinson's Disease », *Movement disorders clinical practice*, vol. 5, *4*, pp. 417–426, 2018.

[167] R. R. Shamir, T. Dolber, A. M. Noecker, A. M. Frankemolle, B. L. Walter, and C. C. McIntyre, « A method for predicting the outcomes of combined pharmacologic and deep brain stimulation therapy for Parkinson's disease », *in International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, 2014, pp. 188–195.

[168] M. Peralta, J. S. Baxter, A. R. Khan, C. Haegelen, and P. Jannin, « Striatal shape alteration as a staging biomarker for Parkinson's Disease », *NeuroImage: Clinical*, vol. 27, 2020.

[169] M. Peralta, P. Jannin, C. Haegelen, and J. Baxter, « Data Imputation and Compression For Parkinson's Disease Clinical Questionnaires (submitted) », 2020.

[170] C. C. McIntyre, S. Miocinovic, and C. R. Butson, « Computational analysis of deep brain stimulation », *Expert review of medical devices*, vol. 4, *5*, pp. 615–622, 2007.

[171] K. A. Nestor, J. D. Jones, C. R. Butson, T. Morishita, C. E. Jacobson IV, D. A. Peace, D. Chen, K. D. Foote, and M. S. Okun, « Coordinate-based lead location does not predict Parkinson's disease deep brain stimulation outcome », *PloS one*, vol. 9, *4*, 2014.

[172] R. Verhagen, L. J. Bour, V. J. Odekerken, P. van den Munckhof, P. R. Schuurman, and R. de Bie, « Electrode Location in a Microelectrode Recording-Based Model of the Subthalamic Nucleus Can Predict Motor Improvement After Deep Brain Stimulation for Parkinson's Disease », *Brain sciences*, vol. 9, *3*, p. 51, 2019.

[173] I. Aviles-Olmos, Z. Kefalopoulou, E. Tripoliti, J. Candelario, H. Akram, I. Martinez-Torres, M. Jahanshahi, T. Foltynie, M. Hariz, L. Zrinzo, *et al.*, « Long-term outcome of subthalamic nucleus deep brain stimulation for Parkinson's disease using an MRI-guided and MRI-verified approach », *J Neurol Neurosurg Psychiatry*, vol. 85, *12*, pp. 1419–1425, 2014.

[174] A. Horn, M. Reich, J. Vorwerk, N. Li, G. Wenzel, Q. Fang, T. Schmitz-Hübsch, R. Nickl, A. Kupsch, J. Volkmann, *et al.*, « Connectivity predicts deep brain stimulation outcome in P arkinson disease », *Annals of neurology*, vol. 82, *1*, pp. 67–78, 2017.

[175] S.-C. Park, J. K. Lee, S. M. Kim, E. J. Choi, and C. S. Lee, « Systematic stereotactic error reduction using a calibration technique in single-brain-pass and multitrack deep brain stimulations », *Operative Neurosurgery*, vol. 15, *1*, pp. 72–80, 2018.

[176] C. S. Lozano, M. Ranjan, A. Boutet, D. S. Xu, W. Kucharczyk, A. Fasano, and A. M. Lozano, « Imaging alone versus microelectrode recording–guided targeting of the STN in patients with Parkinson's disease », *Journal of neurosurgery*, vol. 130, *6*, pp. 1847–1852, 2018.

[177] P. S. Lee, G. M. Weiner, D. Corson, J. Kappel, Y.-F. Chang, V. R. Suski, S. B. Berman, H. Homayoun, A. D. Van Laar, D. J. Crammond, *et al.*, « Outcomes of

interventional-MRI versus microelectrode recording-guided subthalamic deep brain stimulation », *Frontiers in neurology*, vol. 9, p. 241, 2018.

[178] X. Liu, J. Zhang, K. Fu, R. Gong, J. Chen, and J. Zhang, « Microelectrode recording–guided versus intraoperative magnetic resonance imaging–guided subthalamic nucleus deep brain stimulation surgery for Parkinson disease: a 1-year follow-up study », *World neurosurgery*, vol. 107, pp. 900–905, 2017.

[179] K. R. Wan, T. Maszczyk, A. A. Q. See, J. Dauwels, and N. K. K. King, « A review on microelectrode recording selection of features for machine learning in deep brain stimulation surgery for Parkinson's disease », *Clinical Neurophysiology*, vol. 130, *1*, pp. 145–154, 2019.

[180] A. Moran, I. Bar-Gad, H. Bergman, and Z. Israel, « Real-time refinement of subthalamic nucleus targeting using Bayesian decision-making on the root mean square measure », *Movement disorders: official journal of the Movement Disorder Society*, vol. 21, *9*, pp. 1425–1431, 2006.

[181] A. Zaidel, A. Spivak, L. Shpigelman, H. Bergman, and Z. Israel, « Delimiting subterritories of the human subthalamic nucleus by means of microelectrode recordings and a Hidden Markov Model », *Movement disorders*, vol. 24, *12*, pp. 1785–1793, 2009.

[182] H. Cagnan, K. Dolan, X. He, M. F. Contarino, R. Schuurman, P. van den Munckhof, W. J. Wadman, L. Bour, and H. C. Martens, « Automatic subthalamic nucleus detection from microelectrode recordings based on noise level and neuronal activity », *Journal of neural engineering*, vol. 8, *4*, p. 046 006, 2011.

[183] W. Chaovalitwongse, Y. Jeong, M. K. Jeong, S. Danish, and S. Wong, « Pattern recognition approaches for identifying subcortical targets during deep brain stimulation surgery », *IEEE intelligent systems*, vol. 26, *5*, pp. 54–63, 2011.

[184] H. D. V. Cardona, J. B. Padilla, R. Arango, H. Carmona, M. A. Alvarez, E. G. Estellés, and A. A. Orozco, « NEUROZONE: On-line recognition of brain structures in stereotactic surgery-application to Parkinson's disease », *in 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, IEEE, 2012, pp. 2219–2222.

[185] K. Ciecierski, T. Mandat, R. Rola, Z. W. Raś, and A. W. Przybyszewski, « Computer aided subthalamic nucleus (STN) localization during deep brain stimulation (DBS) surgery in Parkinson's patients », *in Annales Academiae Medicae Silesiensis*, vol. 5, 2014, pp. 275–283.

[186] L. Schiaffino, A. R. Muñoz, J. G. Martınez, J. F. Villora, A. Gutiérrez, I. M. Torres, *et al.*, « STN area detection using K-NN classifiers for MER recordings in Parkinson patients during neurostimulator implant surgery », *in Journal of Physics: Conference Series*, IOP Publishing, vol. 705, 2016, p. 012 050.

[187] P. Karthick, K. R. Wan, R. Yuvaraj, A. A. See, N. K. K. King, and J. Dauwels, « Detection of Subthalamic Nucleus using Time-Frequency Features of Microelectrode recordings and Random Forest Classifier », *in 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, IEEE, 2019, pp. 4164–4167.

[188] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, *et al.*, « Gradient-based learning applied to document recognition », *Proceedings of the IEEE*, vol. 86, *11*, pp. 2278–2324, 1998.

[189] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, « Aggregated residual transformations for deep neural networks », *in Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[190] S. Gross and M. Wilber, « Training and investigating residual nets », *Facebook AI Research*, 2016.

[191] Y. Kim and T. Moon, « Human detection and activity classification based on micro-Doppler signatures using deep convolutional neural networks », *IEEE geoscience and remote sensing letters*, vol. 13, *1*, pp. 8–12, 2015.

[192] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, *et al.*, « CNN architectures for large-scale audio classification », *in 2017 ieee international conference on acoustics, speech and signal processing (icassp)*, IEEE, 2017, pp. 131–135.

[193] Y. M. Costa, L. S. Oliveira, and C. N. Silla Jr, « An evaluation of convolutional neural networks for music classification using spectrograms », *Applied soft computing*, vol. 52, pp. 28–38, 2017.

[194] F. Mamalet and C. Garcia, « Simplifying convnets for fast learning », *in International Conference on Artificial Neural Networks*, Springer, 2012, pp. 58–65.

[195] T. Martin, M. Peralta, G. Gilmore, P. Sauleau, C. Haegelen, P. Jannin, and J. Baxter, « Extending Convolutional Neural Networks for Localizing the Subthalamic Nucleus from Micro-Electrode Recordings in Parkinson's Disease », *IEEE Transactions on Biomedical Engineering (submitted)*, 2020.

[196] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, « A survey on deep transfer learning », *in International conference on artificial neural networks*, Springer, 2018, pp. 270–279.

[197] Y. Zhang and Q. Yang, « An overview of multi-task learning », *National Science Review*, vol. 5, *1*, pp. 30–43, 2018.

[198] P.-F. D'Haese, P. E. Konrad, and B. M. Dawant, « Big Data and Deep Brain Stimulation », *in Neuromodulation*, Elsevier, 2018, pp. 137–145.

[199] J. Baxter, M. Peralta, and P. Jannin, « Validating Medical Information Processing Algorithms in the Age of Machine Learning », *International Journal of Computer Assisted Radiology and Surgery (submitted)*, 2020.

[200] M. Khosravi, S. F. Atashzar, G. Gilmore, M. S. Jog, and R. V. Patel, « Intraoperative Localization of STN During DBS Surgery Using a Data-Driven Model », *IEEE Journal of Translational Engineering in Health and Medicine*, vol. 8, pp. 1–9, 2020.

[201] S. Sivaranjini and C. Sujatha, « Deep learning based diagnosis of Parkinson's disease using convolutional neural network », *Multimedia Tools and Applications*, pp. 1–13, 2019.

[202] G. Vandewiele, I. Dehaene, G. Kovács, L. Sterckx, O. Janssens, F. Ongenae, F. De Backere, F. De Turck, K. Roelens, J. Decruyenaere, *et al.*, « Overly Optimistic Prediction Results on Imbalanced Data: Flaws and Benefits of Applying Oversampling », *arXiv preprint arXiv:2001.06296*, 2020.

[203] P. Fergus, P. Cheung, A. Hussain, D. Al-Jumeily, C. Dobbins, and S. Iram, « Prediction of preterm deliveries from EHG signals using machine learning », *PloS one*, vol. 8, *10*, e77154, 2013.

[204] V. Cheplygina, M. de Bruijne, and J. P. Pluim, « Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis », *Medical image analysis*, vol. 54, pp. 280–296, 2019.

[205] N. Rieke, J. Hancox, W. Li, F. Milletari, H. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. Landman, K. Maier-Hein, *et al.*, « The future of digital health with federated learning », *arXiv preprint arXiv:2003.08119*, 2020.

[206] M. M. McCormick, X. Liu, L. Ibanez, J. Jomier, and C. Marion, « ITK: enabling reproducible research and open science », *Frontiers in neuroinformatics*, vol. 8, p. 13, 2014.

[207] R. Kikinis, S. D. Pieper, and K. G. Vosburgh, « 3D Slicer: a platform for subject-specific image analysis, visualization, and clinical support », *in Intraoperative imaging and image-guided therapy*, Springer, 2014, pp. 277–289.

[208] A. Horn, N. Li, T. A. Dembek, A. Kappel, C. Boulay, S. Ewert, A. Tietze, A. Husch, T. Perera, W.-J. Neumann, *et al.*, « Lead-DBS v2: Towards a comprehensive pipeline for deep brain stimulation imaging », *Neuroimage*, vol. 184, pp. 293–316, 2019.

[209] E. Gibson, W. Li, C. Sudre, L. Fidon, D. I. Shakir, G. Wang, Z. Eaton-Rosen, R. Gray, T. Doel, Y. Hu, *et al.*, « NiftyNet: a deep-learning platform for medical imaging », *Computer methods and programs in biomedicine*, vol. 158, pp. 113–122, 2018.

[210] K. J. Gorgolewski, T. Auer, V. D. Calhoun, R. C. Craddock, S. Das, E. P. Duff, G. Flandin, S. S. Ghosh, T. Glatard, Y. O. Halchenko, *et al.*, « The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments », *Scientific data*, vol. 3, *1*, pp. 1–9, 2016.

[211] T. Kapur, S. Pieper, A. Fedorov, J.-C. Fillion-Robin, M. Halle, L. O'Donnell, A. Lasso, T. Ungi, C. Pinter, J. Finet, *et al.*, *Increasing the impact of medical image computing using community-based open-access hackathons: The NA-MIC and 3D Slicer experience*, 2016.

[212] E. Strubell, A. Ganesh, and A. McCallum, « Energy and policy considerations for deep learning in NLP », *arXiv preprint arXiv:1906.02243*, 2019.

[213] D. R. So, C. Liang, and Q. V. Le, « The evolved transformer », *arXiv preprint arXiv:1901.11117*, 2019.

[214] A. Moro and L. Lonza, « Electricity carbon intensity in European Member States: Impacts on GHG emissions of electric vehicles », *Transportation Research Part D: Transport and Environment*, vol. 64, pp. 5–14, 2018.

[215] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, « Quantifying the carbon emissions of machine learning », *arXiv preprint arXiv:1910.09700*, 2019.

[216] P. Henderson, J. Hu, J. Romoff, E. Brunskill, D. Jurafsky, and J. Pineau, « Towards the Systematic Reporting of the Energy and Carbon Footprints of Machine Learning », *arXiv preprint arXiv:2002.05651*, 2020.

[217] D. Banta, « What is technology assessment? », *International journal of technology assessment in health care*, vol. 25, *S1*, pp. 7–9, 2009.

[218] P. Jannin and W. Korb, « Assessment of image-guided interventions », *in Image-Guided Interventions*, Springer, 2008, pp. 531–549.

[219] K. Ren, T. Zheng, Z. Qin, and X. Liu, « Adversarial Attacks and Defenses in Deep Learning », *Engineering*, 2020.

[220] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, « A survey on bias and fairness in machine learning », *arXiv preprint arXiv:1908.09635*, 2019.

[221] M. J. Rigby, « Ethical dimensions of using artificial intelligence in health care », *AMA Journal of Ethics*, vol. 21, *2*, pp. 121–124, 2019.

[222] H. R. Sullivan and S. J. Schweikart, « Are current tort liability doctrines adequate for addressing injury caused by AI? », *AMA journal of ethics*, vol. 21, *2*, pp. 160–166, 2019.

[223] M. Fenech, N. Strukelj, and O. Buston, « Ethical, social, and political challenges of artificial intelligence in health », *London: Wellcome Trust Future Advocacy*, 2018.

[224] A. Diot, M. Peralta, J. Baxter, P. Jannin, and C. Haegelen, « Exploring the acceptability of PassFlow, an AI based decision support software for Deep Brain Stimulation », *Stereotactic and Functional Neurosurgery (submitted)*, 2020.

[225] F. Hell, C. Palleis, J. H. Mehrkens, T. Koeglsperger, and K. Bötzel, « Deep brain stimulation programming 2.0: future perspectives for target identification and adaptive closed loop stimulation », *Frontiers in neurology*, vol. 10, p. 314, 2019.

[226] A. A. Kühn and J. Volkmann, « Innovations in deep brain stimulation methodology », *Movement Disorders*, vol. 32, *1*, pp. 11–19, 2017.

[227] W.-J. Neumann, R. S. Turner, B. Blankertz, T. Mitchell, A. A. Kühn, and R. M. Richardson, « Toward electrophysiology-based intelligent adaptive deep brain stimulation for movement disorders », *Neurotherapeutics*, vol. 16, *1*, pp. 105–118, 2019.

UNIVERSITÉ DE RENNES 1

Titre : Méthodes pilotées par les données pour aider la décision clinique pour la Stimulation Cérébrale Profonde pour la Maladie de Parkinson

Mot clés : Stimulation Cérébrale Profonde, Maladie de Parkinson, Apprentissage Machine, Système d'Aide à la Décision Clinique

Résumé : La Stimulation Cérébrale Profonde (SCP) est une thérapie efficace pour traiter les maladies des mouvements anormaux, telle que la Maladie de Parkinson (MP). Le succès de la SCP dépend de nombreuses variables issues d'un grand nombre de modalités de données. Divers problèmes sont rencontrés tout au long de la prise en charge du patient, de sa sélection à la procédure elle-même et au suivi post-opératoire, dénotant un besoin urgent de développer des outils d'assistance informatique.

Dans cette thèse, nous proposons deux systèmes, basés sur l'apprentissage machine, afin de résoudre deux problèmes cliniques concrets. Premièrement, nous proposons un outil capable d'aider les cliniciens dans le choix de sélection des patients et des cibles de stimulation. Notre méthode est capable de prédire les résultats cliniques (moteurs, neuropsychologiques, cognitifs, etc.) de la SCP à partir de biomarqueurs multimodaux préopératoires. Deuxièmement, nous proposons un outil permettant d'accélérer grandement la chirurgie en assistant la localisation du noyau cible via un traitement en temps réel du signal électrophysiologique provenant du cerveau du patient, à partir d'enregistrements par micro-électrodes d'une seconde seulement.

Title: Data driven methods to support decision making in Deep Brain Stimulation for Parkinson's Disease

Keywords: Deep Brain Stimulation, Parkinson's Disease, Machine Learning, Clinical Decision Support System

Abstract: Deep Brain Stimulation (DBS) is a successful and encouraging way of treating abnormal movement diseases, such as Parkinson's Disease (PD). The success of the surgical procedure depends on many variables, most of which are derivative from a great number of modalities. Various problems gravitate throughout the care of the patient, from its screening, to the procedure itself and the stimulation follow-up, creating an urging need to develop computer assisting tools.

In this thesis, we used data-driven methods to design two systems in order to address two concrete clinical applications. Firstly, we propose a tool to assist clinicians in decision making for selecting patients and stimulation targets. It consists in a data-driven method which is able to predict the clinical outcomes (motor, neuropsychologic, cognitive etc.) of the surgery, from pre-operative multimodal biomarkers. Secondarily, we propose to greatly fasten the surgical procedure by automatizing the location of the target nucleus via a real time treatment of the electrophysiological signal arising from the patient's brain, from micro-electrode recordings (MER). Our method is able, in one second, to accurately analyse the MER and predict whether the electrode lead is inside the STN or not, and does not require any parameter tuning nor calibration to work on a new data source.