

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

**Prognosis for Degenerative Cervical Myelopathy: a Computer Learning
Approach on the AOspine Database**

LUCAS ROUHIER

Institut de génie biomédical

Mémoire présenté en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
Génie biomédical

Août 2020

POLYTECHNIQUE MONTRÉAL
affiliée à l'Université de Montréal

Ce mémoire intitulé :

**Prognosis for Degenerative Cervical Myelopathy: a Computer Learning
Approach on the AOspine Database**

présenté par **Lucas ROUHIER**

en vue de l'obtention du diplôme de *Maîtrise ès sciences appliquées*
a été dûment accepté par le jury d'examen constitué de :

Nikola STIKOV, président

Julien COHEN-ADAD, membre et directeur de recherche

Samuel KADOURY, membre

DEDICATION

*À mes parents Francis et Martine
à mes amis qui m'ont soutenus*

...

ACKNOWLEDGEMENTS

I would like to warmly thank my supervisor Julien Cohen-Adad giving me the opportunity to work on this study the last two years. The suggestions, discussion and meetings he offered to help me work on this study, his approach on research and his patience during this master's project were greatly appreciated. He allowed me to understand the importance of the project as well as many critical aspect about leading a research project in biomedical science.

I would like to also thank all the members of NeuroPoly Lab for fruitful discussions and valuable suggestions, especially Charley Gros, Anthime Bucquet, Olivier Vincent, Andréanne Lemay, Laura Duval and Christian Perone for their inputs on the work, as well as Alexandru Foias who helped with everything in the lab.

I would like to thank Drs. Michael Fehlings and Michael Weber for sharing the AOspine database, and Dr. Akbar for his input on the work and help with the database.

We thank the NVIDIA Corporation for the donation of a GPU.

RÉSUMÉ

Description du problème : La myélopathie cervicale dégénérative (MCD) est une condition particulière liée à l'âge touchant environ 600 adultes par million à travers le monde [1]. Elle résulte d'une compression spontanée de la moelle épinière, causée par une excroissance osseuse d'une vertèbre cervicale ou bien d'un des disques intervertébraux. Dans les deux cas, la moelle se retrouve comprimée et le patient commence à perdre sensations et contrôle moteur. Cette maladie a un très fort impact socio-économique. En effet, les patients perdent graduellement l'usage de leurs membres, les empêchant de vivre et de travailler au fur et à mesure que la compression augmente [2] [3].

Cette maladie est principalement diagnostiquée à l'aide d'évaluations cliniques du contrôle moteur et l'utilisation d'imagerie par résonance magnétique. Le problème reste toutefois complexe, car la compression peut être asymptomatique, et de faibles compressions sont encore difficiles à détecter avec les méthodes d'imagerie actuelle. Certains patients peuvent attendre de long mois avant d'avoir un diagnostic [4].

Fort heureusement, une méthode existe afin de limiter la future perte de contrôle moteur : il s'agit de la chirurgie décompressive. Cette chirurgie peut prendre plusieurs formes et est décidée au cas par cas. Il n'existe pas encore de consensus sur les détails et les approches de telles opérations. Toutefois, cette opération compte de très nombreux risques (p. ex., paralysie C5) [3]. La moelle épinière est en effet une zone sensible, et le chirurgien doit aller travailler au plus proche de cette autoroute nerveuse vitale. De nombreuses complications peuvent suivre. Bien que le résultat possible semble excéder les risques, la réalité est bien différente, car seulement 40 % des opérations ont un réel impact sur le rétablissement des patients. Le pronostic postopératoire est une tâche compliquée et l'opération est donc choisie par défaut. La question demeure sur la possibilité d'établir ce pronostic de manière plus précise. Il n'existe pour l'instant que peu d'étude se penchant sur l'analyse quantitative de données cliniques des patients afin d'obtenir un pronostic postopératoire. L'une des pistes d'exploration serait d'utiliser des méthodes d'intelligence artificielle afin de créer un modèle capable d'aider les chirurgiens dans leur prise de décision. Cela passe par l'exploitation de données cliniques et IRM. Durant les années précédentes, seulement deux études sont apparues exploitant des données similaires dans un but identique. Ces études sont principalement centrées sur l'analyse des données cliniques [5] [6].

Objectifs : L'objectif de cette étude est de créer ou d'exploiter un modèle existant afin de vérifier si l'intelligence artificielle pourrait apporter des solutions à ce problème. Parmi les sous-objectifs de ce travail, le but est notamment de vérifier les hypothèses suivantes :

- L'exploitation d'IRM conjointement avec les données cliniques apporte de meilleurs résultats
- Il est possible d'établir un modèle de pronostic postopératoire pour la myélopathie cervicale dégénérative.

Méthodes et matériel : Nous avons à notre disposition une base donnée de 759 sujets pour lesquels nous avons 135 points de données cliniques ainsi que, pour une partie des sujets, des images IRM de modalité T2 et T1 avec les vues axiales et sagittales. Ces informations cliniques regroupent différentes données médicales et courantes sur le patient telles que l'âge, le sexe, etc... Ces données ont été acquises prospectivement durant une précédente étude : AOSpine [7]. Ces données proviennent de différents centres et offrent donc une bonne hétérogénéité au niveau du contraste des images, simulant correctement des données réelles. Cela est important pour évaluer la capacité de généralisation du modèle. Ces patients souffrent tous de MCD et ont été opérés dans les différents centres. Parmi ces données, nous avons différents scores d'évaluation de leur capacité de contrôle moteur ainsi que de leur sensation (MJOa, SF6D, ...). Ces scores cliniques ont été établis avant l'opération ainsi que 6, 12 et 24 mois après. La différence entre le score préopératoire et le score postopératoire servira de cible. Le score le plus important semble être celui établi sur l'échelle de la modified japanese orthopedic association (MJOa). La différence entre le score préopératoire et le score postopératoire pourra être classifiée en 2 catégories selon la différence minimale significative qui est de 2 points [8]. Une augmentation du score de 2 traduirait donc une amélioration de l'état des patients après chirurgie.

Les données contiennent également des informations sur l'opération subie par le patient qui ne sont a priori pas disponibles dans le cadre de pronostic préopératoire. Toutefois, cela pourrait être utile pour étudier l'impact de ces données sur les performances de notre modèle prédictif.

Les données cliniques ont tout d'abord été manuellement analysées afin d'essayer d'en extraire uniquement les données importantes ainsi que de retirer les données postopératoires dans un premier temps. Cela a permis d'établir un score de base sur un modèle classique type «random forest» fourni dans le package «scikit-learn» de python. Plusieurs modèles de machine learning ont alors été testés pour établir un score maximum atteignable avec ces données. Nous avons également ajouté les données opératoires dans nos modèles afin d'évaluer leur impact sur les performances du modèle .

Par la suite, différents modèles de réseaux de neurones artificiels, principalement convolutionnels, ont été créés afin d'effectuer l'analyse automatique des images IRM. Ces images n'étaient pas les images originales, mais le résultat d'un prétraitement à l'aide de la «spinal cord toolbox» [9]. Différents modèles furent établis : le premier utilisait uniquement les IRM T2 sagittal, le suivant les IRM T2 et T1 sagittal, et le dernier fonctionnait les données T1 et T2 sagittales ainsi que les données cliniques. La partie du réseau traitant les données cliniques fut également testée seule pour vérifier ses performances face au modèle de machine learning établi à l'étape précédente.

Le pipeline donne une précision de prédiction de 72,5 % (soit une amélioration de 8 % par rapport à la baseline) avec une aire sous la courbe (ASC) de 0,73 pour le modèle basé uniquement sur des données cliniques. Toutefois, cela dépend fortement de la quantité de données disponibles. Les modèles d'apprentissage profond ont tendance à overfit ou underfit les données montrant un manque de généralisabilité du modèle, ce qui pourrait s'expliquer par le nombre réduit d'IRM disponibles. L'ajout des données extraites semble fournir au modèle une plus grande capacité puisque l'amélioration par rapport à la baseline atteint 8 % avec une précision de 65,2 % et une ASC de 0,69 avec moins de sujets que le premier modèle.

ABSTRACT

Description of the problem: Spinal injuries may impair patients' motor control as the spinal cord represents a nervous highway connecting the brain and the limb. Some of these injuries arise from accidents others occur progressively; such includes Degenerative Cervical Myelopathy (DCM). Cervical myelopathy is caused by the compression of the spinal cord by an outgrowth of a vertebral body or intervertebral disc, yielding symptoms such as sensorimotor dysfunction or pain. Cervical myelopathy is degenerative, which implies that it only gets worse as time goes by, and the compression increases. This condition is a cause of surgery among 40 adults per million yearly [1].

The diagnosis for this condition is made possible through clinical motor skill testing and magnetic resonance imaging (MRI). However, diagnosis is still a complex problem as the compression can be asymptomatic or, in some cases, not easily visible in MR images at its early stages. The diagnosis can take months, if not years, for some patients. [4]

The current study proposes a decompressive surgery which aims at removing the object causing the compression, or at least, part of it. The goal here is to avoid further compression of the spinal cord and alleviate the existing ones. This operation can take various forms (anterior or posterior approach, bone graft, bone fusion) and includes many risks (e.g., C5 palsy) [3] as this touches the spinal cord, that is, highly sensitive and important to the body. The details of each surgery are then determined case by case by the surgeon as no consensus on the aspects of such operation exists [10]. This exposes a complicated scenario in predicting the outcome of the surgery. Even though the benefits seem to outweigh the risk, the operation is only successful in 35 % [11] of the cases. This success rate is based on the improvement of the patients' sensorimotor skills. As it is complex to predict the outcome, surgery is often seen as a default option; however, there is an open question about the possibility of predicting the outcome of the surgery. To the best of my knowledge, only a handful of studies that utilize quantitative clinical data analysis in predicting post-surgery prognosis exist. One of the leading techniques is the use of data science and artificial intelligence to design a model that will be able to establish this prognosis and assist surgeons in the decision process with AO spine [7] clinical data. The first study exploring this concept was published in 2019 [5], and, since then, only two studies have been conducted to improve on the methodology [6]. These studies mainly focus on the analysis of clinical data.

Objectives: Primary Goal: The primary goal of this study is to develop a model based on artificial intelligence (AI) that can be used in patient prognosis and treatment of DCM

The sub-objectives for the study are:

- To establish the efficacy of using MRI in conjunction with clinical data in providing better results in treatment and prognosis.
- To establish a postoperative prognostic model for DCM.

Material and method: The database consists of 769 subjects, most of whom have T2 and T1 MRI images with an axial and sagittal view and 135 clinical data points. These data was acquired prospectively during the AOSpine study. These data came from numerous centers and offer a functional heterogeneity in image contrast, making it close to a real-life scenario. The inclusion criteria for patients used as participants in this study were to be suffering from DCM and have undergone surgery. Among these data, we have different clinical evaluations of their sensorimotor capacities according to various scales (MJOa, SF6D. . .). These scores were established before the operation as well as 6, 12, and 24 months after the surgery. The goal is to predict the difference between pre-surgery and post-surgery scores. We evaluated which score is the most relevant among the three measures at 6, 12, and 24 months. This target can be classified into two categories according to the minimum clinically significant difference [8], which is two. The Modified Japanese Orthopedic Association (MJOA) has been used in this study to record and gauge the results. An increase of 2 points or more would, therefore, reflect an improvement in the condition of the patient after the surgery. The data also contain information from the surgery performed on the patient, which isn't available in the preoperative prognosis. However, this may be useful to study the impact of these data on the performance of our predictive model.

Three approaches were tested in order to try to improve on the current prognosis. The first one aims at exploiting clinical data, the second one aimed at leveraging deep learning method to use images as well as clinical data as input, the third one exploit features extracted from Magnetic Resonance (MR) images.

Clinical data was processed through machine learning. This was done after a pre-processing step aimed at removing the non-relevant features in order to get the best results possible. To process available MRI data jointly with clinical data, two different strategies were implemented. The first one is geared towards the use of deep learning and the exploitation of artificial neural networks, which were fed pre-processed sagittal images and clinical data. The models used were a Resnet and a custom model to use both T1w and T2w MR images. The model is based on three different modules. Two of them are similar and were used

to process and encode features from the image. They were created with convolutional filter and inception modules. The second one relies on feature extraction from the axial images through a semi-automatic processing pipeline. These features were then added to the existing clinical one to improve the ability of the model to generalize to the unseen cases. All models were tested for accuracy on unseen data. Data were split between training, validation, and testing (80%,10%,10%, respectively). Results show an accuracy of 72.5 % (8 % improvement from the baseline) with an area under the curve (AUC) of 0.75 for the model-based solely on clinical data. This is, however, heavily dependent on the quantity of available data. Deep learning models tend to overfit or underfit the data showing a lack of generalizability from the model, which could be explained by the reduced number of available MRI. The added extracted feature seems to provide the model with valuable insight as the improvement from the baseline reaches 8 % with an accuracy of 65.2% and an AUC of 0.68 with less subject than the first model.

TABLE OF CONTENTS

DEDICATION	iii
ACKNOWLEDGEMENTS	iv
RÉSUMÉ	v
ABSTRACT	viii
TABLE OF CONTENTS	xi
LIST OF FIGURES	xiv
LIST OF SYMBOLS AND ACRONYMS	xvi
CHAPTER 1 INTRODUCTION	1
1.1 Problem Statement	2
1.2 Research Objectives	3
1.3 Thesis Outlines	3
CHAPTER 2 LITERATURE REVIEW	4
2.1 Spinal Cord	4
2.1.1 Nervous System	4
2.1.2 Spinal Cord Tracts and Network	6
2.1.3 Spinal Canal	8
2.1.4 Spinal Cord Anatomy	9
2.2 MRI	10
2.2.1 MRI basics	10
2.2.2 MRI Contrast and Orientation	11
2.2.3 Challenges of Spinal MRI	13
2.3 Spinal cord lesion and compression: Degenerative cervical Myelopathy	13
2.3.1 Pathophysiology	13
2.3.2 Impact and Diagnosis	14
2.3.3 Current Solution and Issues	15
2.4 Biomedical Data Processing Through Machine Learning	16
2.4.1 Machine Learning Basics	16

2.4.2	Deep Learning	17
2.4.3	Application in Biomedical Imaging	19
2.4.4	Machine Learning Processing of DCM	21
2.5	Contributions	22
2.6	Spinal Cord Toolbox	22
2.7	Vertebral Labeling by Deep Learning	23
2.8	Prognosis	24
CHAPTER 3 MATERIAL AND METHOD		25
3.1	Material: Dataset presentation	25
3.1.1	Dataset Constitution and Specific	25
3.1.2	Patients and Feature Selection	26
3.1.3	Dataset Statistics	28
3.1.4	Patient's Evaluation and Outcome	30
3.2	Approach of the Problem	30
3.3	Image Preprocessing	31
3.3.1	Spinal Cord Segmentation	31
3.3.2	Spinal Cord Straightening	31
3.3.3	Image Cropping	32
3.3.4	Inter-vertebral Disc Labeling	33
3.3.5	Feature Engineering	33
3.4	Model and Optimization	35
3.4.1	Machine Learning	35
3.4.2	Regularization & Overfitting	37
3.4.3	Deep Learning Architecture	37
3.4.4	Training and Model Selection	40
3.4.5	Machine Learning Model Selection	40
3.4.6	Optimization of the Problem	42
3.4.7	Optimization of Models	43
3.5	Evaluation and Metrics	43
3.5.1	Metrics	43
3.5.2	Evaluation	44
CHAPTER 4 RESULTS		45
4.0.1	Model Selection	45
4.0.2	Surgery Details	45
4.0.3	Bayesian Optimization	47

4.0.4	Ratio and MCID	49
4.0.5	Deep Learning	50
4.0.6	Atlas Analysis	51
CHAPTER 5 CONCLUSION AND RECOMMENDATION		53
5.1	Summary of Works	53
5.2	Future Research	53
5.2.1	Surgery Decision	53
5.2.2	DWI	53
REFERENCES		55

LIST OF FIGURES

2.1	Figure showing the structure of a neuron.	5
2.2	White matter tracts in a slice of spinal cord	7
2.3	Spinal levels representation along the vertebral column	9
2.4	MRI basics physics	11
2.5	Common MR contrast	12
2.6	Various causes of DCM. CSF = Cerebro spinal fluid. PPL = posterior longitudinal ligament	14
2.7	Model with its receptive fields (40x40) on the input. [12]	23
3.1	Example MR images from the studied dataset	26
3.2	Dataset division	28
3.3	Age distribution	29
3.4	Application of the sct_deepseg method on a T2 sagittal image from the AOSpine dataset	32
3.5	Straightening example	32
3.6	Cropping example	33
3.7	labeling example	33
3.8	Compression example	34
3.9	Example of the atlas registration	35
3.10	Random forest visualization with three trees	36
3.11	Example of skip connexion in a deep neural network	38
3.12	inception module	40
3.13	example of 4 folds cross validation	41
4.1	<i>boxplot showing cross-validation scores without operation data</i>	45
4.2	<i>Ten-folds cross-validation AUC with operation information</i>	46
4.3	<i>Ten-folds cross-validation AUC without operation information</i>	46
4.4	<i>A) Ten-folds cross-validation AUC without Bayesian optimization using Xgboost. B) Ten-folds cross-validation AUC with Bayesian optimization using the Xgboost model.</i>	47
4.5	<i>A) Ten-fold cross-validation AUC without Bayesian optimization using the Random Forest model. B) Ten-folds cross-validation AUC with Bayesian optimization using the Random Forest model</i>	48
4.6	Ten-folds cross-validation on the dataset with recovery ratio as ground truth.	49
4.7	ResNet-29 accuracy during training process	50

4.8	AUC obtained on the dataset completed with atlas features, using the minimum clinically significant difference as ground truth for the training. A) Results obtained with Random Forest. B) Results obtained with Xgboost	51
-----	---	----

LIST OF SYMBOLS AND ACRONYMS

DCM	Degenerative cervical Myelopathy
AI	Artificial intelligence
DL	Deep Learning
ML	Machine Learning
MJOa	Modified Japanese orthopedic association scale
MRI	Magnetic Resonance Imaging
CSF	Cerebro-Spinal Fluid
SC	Spinal Cord
SCT	Spinal Cord Toolbox
FOV	Field of view
GM	Gray Matter
WM	White Matter
DTI	Diffusion Tensor Imaging
SVM	Support Vector Machine
ANN	Artificial Neural Network
CNS	Central Nervous system
LR	Learning Rate
MR	Magnetic resonance
MRI	Magnetic resonance imaging
FOV	Field of view
SVM	Support Vector Machine
RF	Random Forest
MS	Multiple Sclerosis

CHAPTER 1 INTRODUCTION

Degenerative cervical myelopathy (DCM) is a complex neurodegenerative illness affecting a large number of adults worldwide [2]. Degenerative cervical myelopathy (DCM) is a result of spontaneous compression of the spinal cord at the cervical level, which impairs the patient motor control. As the disease progresses, the motor control decreases toward paraplegia and tetraplegia. It is age-related and quite hard to diagnose [4], thus justifying the occurrence of many severe cases. The dataset used in the study was prospectively acquired during the AOspine study [7] in multiple centers. It presents a good heterogeneity data sample making it a realistic scenario for real-life applications. The patient sensorimotor state is evaluated on the Modified Japanese Orthopedic Association scale (MJOa), which represents its current sensorimotor impairment. The current solution to relieve the patient and try to improve his motor control is to perform decompressive surgery. This operation is aimed at removing the origin of the compression. However, in addition to its risks the operation is only successful in 35% of the cases [11]. Surgery is qualified as successful if there is an improvement of at least two points of MJOa score, which represents the minimum difference clinically significant [8]. Post-surgical prognosis is currently based on compression severity and other clinical factors. However, it holds a very limited prognosis value [13]. Diagnosis and prognosis come from the analysis of MRI images and the evaluation of clinical scores. There are asymptomatic compressions that pose a serious challenge in diagnosis with the current imaging techniques. These result in delayed diagnosis leading to more severe cases. This study aims at developing a prognosis tool using machine learning and deep learning to predict the patient's state after the surgery. Deep learning (DL) and machine learning (ML) models are part of the ongoing artificial intelligence development gaining traction on its applicability in the medical field. The recent improvements in machine learning for medical images and data analysis justify this new approach. Various algorithms and models have been trained and tested. This work presents three different approaches that were tested in order to perform this prognosis. The first approach relies on the exploitation of clinical data with common machine learning models. The second approach aims at using magnetic resonance images as well as clinical data with Deep Learning model, while the third one aims at extracting features from images to add to the clinical data, before using machine learning models.

1.1 Problem Statement

The current problem in Degenerative Cervical Myelopathy (DCM) diagnosis lies in the complexity of disease identification and the poor prognosis in the current methodology. One key issue with the traditional methods is the adverse lack of generalization with traditional methods. The current prognosis, however, is aimed at improving this through the use of data science analysis and the use of AI models.

Patient difference & availability of data: Every patient has a different background that will need to be taken into account, this can lead to a different output; smokers, for instance, have less successful surgery than non-smoking patients, as stated by various researchers and scholars in the field and backed by immense data. The considered cohort group of 769 patients represents a low number of datasets to establish correlation and causation. A small dataset might create a false correlation for real-life applications, thus hindering the performance of the model, especially in the testing phase. Moreover, even though we have more than a hundred points of data for every patient, a relevant feature could still be missing. There is, of course, a risk of bias as some of the criteria might be subject to the investigator’s opinion and personal bias. An excellent example that highlights the difference between patients’ triage and outcomes is the possibility of asymptomatic compression. Scholars suggest that asymptomatic compression can happen with compression levels between 8% and 57% [11]. This vast difference represents the inequality between patients’ diagnosis and treatment, and it represents the first hurdle which the model has to get over. The availability of each type of exploited data is also not guaranteed due to the willingness to use high resolution axial and sagittal images for better feature extraction and in the hope of improving model performance.

Operation difference: The surgery process is subject to variability. For example, the surgery type may vary from case-to-case, as there is no consensus nor standard procedure for such operation. This represents a less predictable factor that our model will need to learn as this is more a case-by-case decision. Depending on the approach, risks of complications are different, which could affect patients’ scores after surgery [14]. Besides the procedure, the patient may present various responses during the surgery, causing more prolonged operations or need for intensive care. Both of these parameters are not easily predictable. The surgery involves a sensitive organ, the spinal cord, which implies that even small issues can have huge impacts. Moreover, the risk of complications that is implied in such surgery is important (e.g., C5 palsy happens in 5.3% of the cases) [15], which mitigates the score at follow-ups and might hinder the model performance. Such a hindrance calls for another pattern that need to be found to predict the complication.

1.2 Research Objectives

The main objective of the study is to create an AI model or exploit an existing one to make a post-surgery prognosis for patients suffering from DCM. Among the sub-objectives, we will try to see the influence of the operation on the model's performance and determine a way to manually extract relevant features from MRI with a newly developed automatic method. This pipeline could be useful for future similar studies.

1.3 Thesis Outlines

This thesis starts with an introductory chapter before a literature review section aimed at reviewing the nervous system and how it is affected by the DCM. We will then describe the imaging methods, which will help in explaining the processing used in this study. It will be followed by the description of the condition (origin, pathophysiology). In concluding this literature review, some useful concepts in deep learning and machine learning will be presented, along with some relevant applications and existing studies on the DCM prognosis issue. After listing the contribution provided to the lab and the community during the research phase, the methodology will be presented by first detailing the dataset and its acquisition mechanism. This will be followed by an explanation of the various preprocessing performed on the images and on the data before finishing with the tested algorithm and selection process. After the presentation of the results, the limitation and plausible next steps for future studies will be discussed before concluding on our work.

CHAPTER 2 LITERATURE REVIEW

This chapter aims at providing information about the spinal cord (2.1), which is the physiological system affected by DCM. Principles of magnetic resonance imaging are then described (2.2), with a focus on spinal cord imaging. The third section presents the fundamental aspects of computer learning as well as its application to important biomedical issues and similar studies.

2.1 Spinal Cord

This section aims at providing an overview of the complex neurologic system; that is, the spinal cord to provide insight into the consequences of spinal cord compression, which will be later discussed. This section briefly presents the Central Nervous system (CNS), the spinal cord tracts, and canal, as well as its morphometry.

2.1.1 Nervous System

The nervous system is composed of two parts: the Central Nervous System (CNS) and the peripheral nervous system (PNS). The CNS is aimed at processing and transporting information and command while the PNS transmits information back and forth between the CNS and the muscles and organs. The CNS consists of the brain and the spinal cord [16]. It is responsible for the processing of all sensory information as well as the impulse for the motor control that exhibits the human body. It also uses the sensory information to trigger some metabolic processes to adjust the system, for example, during an effort (increasing heart rate to increase the amount of oxygen to the right muscles). The nervous system is composed of neurons, needed to transport and emit electrical signals, and are supported by glial cells. [17]

2.1.1.1 Neurons

Neurons are nerve cells consisting of three parts: the axon, the dendrites, and the soma. The axon is a long tail used to propagate an electrical signal. The dendrites are a series of highly branched outgrowth to get information from neighboring neurons or external stimuli. Between these two parts is the body of the cell, which holds the nucleus and the soma. It contains the genetic information, the ribosomes as well as the machinery for protein synthesis. The neuron also has an axon terminal used to transmit information. The basic structure

of a neuron can be seen in figure 2.1. The neuron propagates the information as an electrical signal. Each neuron cell can be myelinated or not, which means that it is enveloped in fat, which helps speed up the electricity propagation.

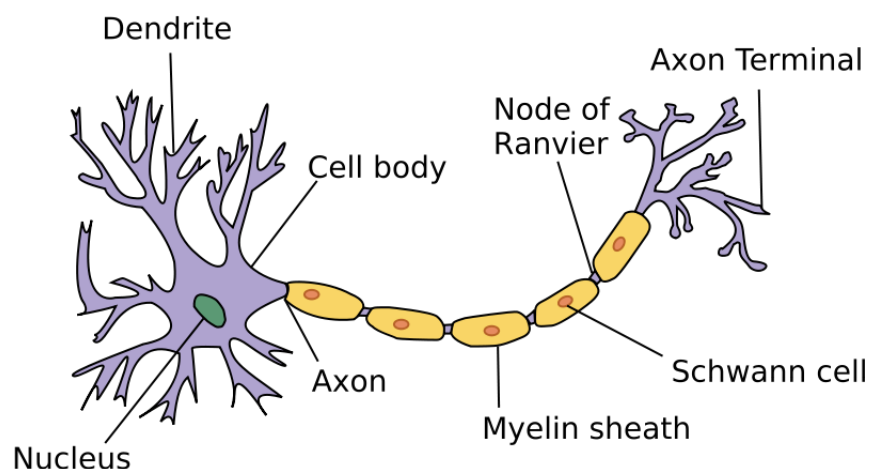


Figure 2.1 Figure showing the structure of a neuron.

source: <https://commons.wikimedia.org/wiki/File:Neuron.svg>

There is a massive amount of neurons in the brain. They can be categorized based on their function, their physiology, and their localization. Neurons can have different shapes that help them in a specific function and are often part of many different neural circuits inside the human body.

Basic propagation: The neurons emit and propagate electrical signals throughout the body. These signals are composed of Action Potential (AP). An AP is an electrical impulse. The creation and transmission of AP are caused by changes in the ionic composition of the intracellular fluid. Transmitted information is coded in action potential frequency. The neurons have a **refractory period**. This represents the minimum time between the creation of two action potential by a single neuron.

At the level of the synapse, which is the gap between the end of one neuron and the beginning of the other, the information is coded in the concentration of neurotransmitters. The pre-synaptic neuron will emit some neurotransmitters that will be caught by the post-synaptic neuron. It triggers a signal from its dendrite to the end of its axon. The neuroreceptor is specific to a particular type of neurotransmitter.

Propagation with Myelin: The speed of the propagation of the action potential alongside the axon is dependent on the nature and the diameter of the axon fiber. Larger fibers offer less resistance to local currents, and therefore, the ion flow will increase and bring adjacent membrane to the activation threshold faster [18].

It also depends on the presence of **myelin** as mentioned in 2.1.1.1. Myelin acts as an insulator, which allows a specific AP to go farther before needing to be regenerated. Even though it can go further, you want to avoid any attenuation. Therefore myelin is not a single sheet of fat all along the axon of one neuron but rather a discontinued sheet of fat. An action potential can only be generated when the myelin is interrupted. Such places along the axon are named nodes of Ranvier. Action potential virtually jumps the myelin coated part of the axon between two nodes of Ranvier. This propagation is much more efficient from a metabolic standpoint and from a speed perspective. Indeed the electrical propagation is faster than the successive adjacent activation. Moreover, this lowers the number of ions exchanged because few action potentials are generated, which reduces metabolic cost.

Even though the brain is the source of the miracle that constitutes the CNS, it would not work without the spinal cord, which is the main artery of transportation for the information required and delivered by the CNS. [17]

2.1.2 Spinal Cord Tracts and Network

The spinal cord is the link between the brain and the rest of the nervous system. The sensory signals, which come from the PNS, come from sensitive receptors in the tissues and transit through the spinal cord before going to the brain through the brainstem. The brain can send numerous commands for motor control to the PNS that transmits them to the effector through the spinal cord. The spinal cord, like the brain, is spatially organized in different parts that have different purposes. Each part is supposed to carry or generate a specific type of signal. In the brain, the various areas are categorized in the cortex (e.g., motor cortex). In the spinal cord, these different areas are divided into multiple tracts or pathways. The two main categories are descending and ascending pathways that are divided into smaller tracts. The ascending tracts bring sensory information from the PNS to the brain, and the descending pathway carries the motor control command from the brain to the motoneurons.

The spinal cord is composed of two different parts: the white matter and the gray matter. The first one is composed of the myelinated axon part of the neuron, and the second one contains the cell bodies of these neurons [19]. These parts are themselves divided into

several tracts that have different origins and destinations.

The gray matter has a butterfly shape, as seen in figure 2.2. The two posterior bumps, called dorsal horns, are mostly composed of sensory nerve cells, while the anterior (ventral) horns mostly host motor nerve cells. The white matter has a defined spatial repartition as well. There are bundles of nerve tracts with specific roles that have specific places. This spatial repartition, as seen in figure 2.2, explains the fact that local spinal cord damage can be associated with different functional impairments.

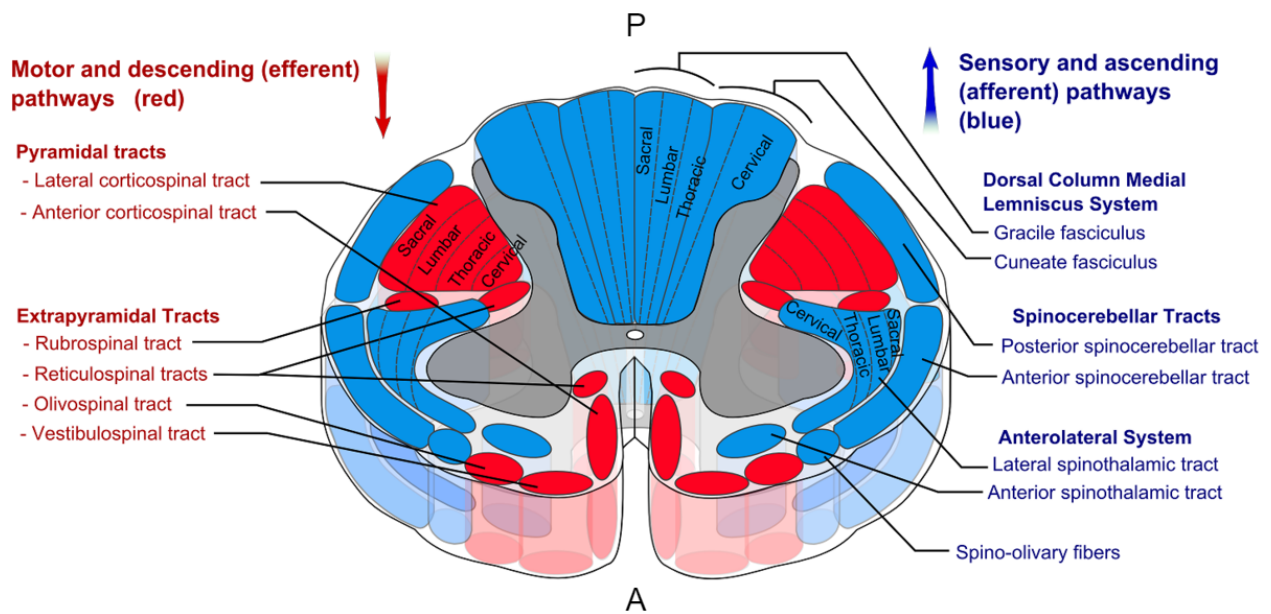


Figure 2.2 White matter tracts in a slice of spinal cord

It shows the various tracts within the spinal cord as well as the representation of white matter (combination of blue and red area) and gray matter (gray area). https://commons.wikimedia.org/wiki/File:Spinal_cord_tracts_-_English.png

There are also spatial differences along the Superior-inferior axis. There is, of course, the diminution of the number of tracts because of the position of muscles and organs (e.g., damage to the lower part of the spine won't affect the arms), but there is also the presence of spinal networks, which will be in a specific location. For example, the Central pattern generator is a spinal network that plays a key role in locomotion [20]. Damages to these types of neural networks can be assimilated to brain damage as this is not only the signal transportation that is lost but the signal generation unit.

The spinal cord is flexible but also needs to be protected, considering its importance that it plays in the life of a human. Therefore it has a particular protection offers by a specific canal described in the next section.

2.1.3 Spinal Canal

The spinal cord is placed between the vertebrae, which are circle-shaped bones. Between each of these vertebrae, there is a cartilaginous disc, which is called **intervertebral** disc. This disc gives the vertebral column its flexibility while the vertebrae give him its sturdiness and provides some protection. The spinal cord is not directly placed inside the vertebrae but floats inside a canal filled with **cerebrospinal fluid** (CSF). The vertebral column is divided into different levels regrouping a certain number of vertebrae: the **cervical** level (c1 to c7), the **thoracic** level (t1 to t12), the **lumbar** level (l1 to l5) and the **sacral** level (s1 to s5). The levels are here listed from top to bottom, so the cervical level is highest one, and the sacral level is located in the lowest part of the spine.

Nerve signals come from the root of the spinal cord and follow some specific path to different muscles and organs, as can be seen in figure 2.3. This shows the spatial repartition mentioned before and explains why lesions have an impact on organs mostly below it.

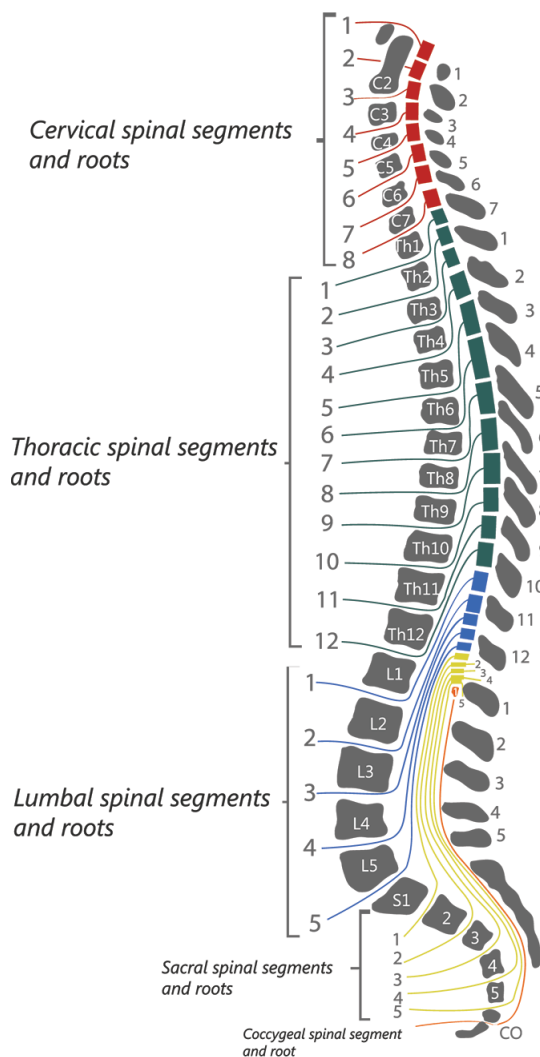


Figure 2.3 Spinal levels representation along the vertebral column

A figure representing the spinal level and the vertebrae level as describe in the previous paragraph. <https://www.sci-info-pages.com/levels-and-classification/>

2.1.4 Spinal Cord Anatomy

The spinal cord is a long tube of neurons that runs through the vertebral column which size evolves with age. It starts in the medulla oblongata, at the base of the brain, and ends up close to the second lumbar vertebrae (represented by L2 in figure 2.3). It possesses an ovoid shape and has a cross-sectional diameter of approximately 1 cm². For the adult human, this represents a length of roughly 45 cm. It is divided into 31 spinal levels that are based on the information they carry. Some of the spinal cord tracts provide information to the upper level of the body and are therefore not found in the lower part of the spinal cord. Their exit/input

possibility determines these segments.

2.2 MRI

2.2.1 MRI basics

Magnetic resonance imaging is used worldwide since its invention in 1977. It is based on the principle of Nuclear Magnetic Resonance (NMR). All nuclei that have an odd number of protons have spin angular momentum, creating a magnetic dipole. The MRI machine creates a strong magnetic field, forcing each spin of the nuclei to be aligned. For the human MRI, this magnetic probation is done by looking at hydrogen protons that are present across all of the human body in the form of water (H₂O). The spin act, in fact, as a magnetic dipole that has a peculiar rotational frequency called the **Larmor frequency**, which is proportional to the strength of the magnetic field.

There are two magnetic fields applied to the patient's body during the imaging: a static one called **B₀** and a rotational one noted **B₁**. The static magnetic field is used to polarize all the angular momentum in the same direction. Its typical strength is between 1.5-7T. During the image acquisition, a stronger magnetic field creates images with greater precision, explaining the continuous high field MRI research. Most current MRI machines use 3 T coils to acquire the images, but 7 T MRI machines are starting to appear as clinical devices. Research devices can create magnetic fields of up to 11 tesla for human applications (or 20+ tesla for animal or NMR systems), allowing for more precise image acquisition.

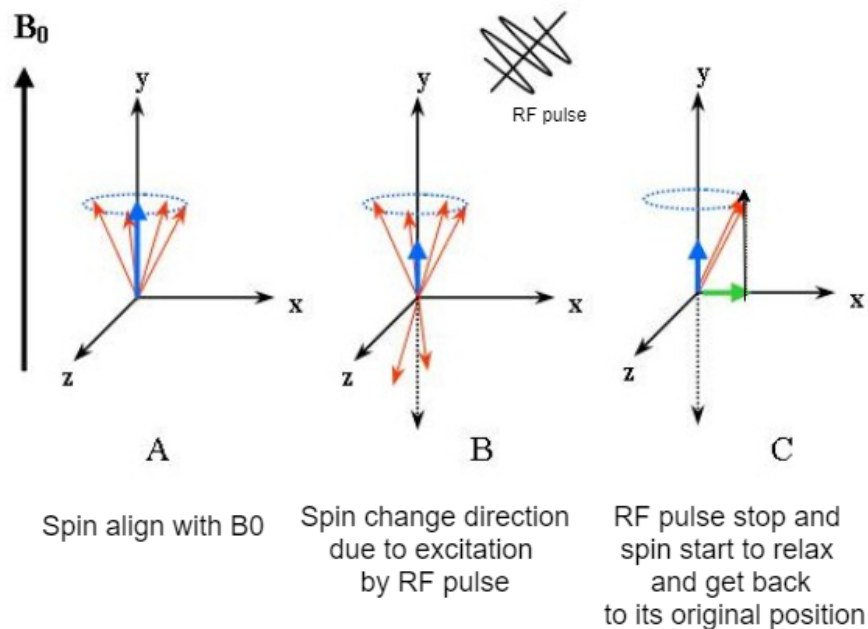


Figure 2.4 MRI basics physics

This shows the basic of MR physics. This relaxation time will help get the T1 characteristic time. image modified from <https://commons.wikimedia.org/wiki/File:Mr4.jpg>

The time it takes for the proton spins to realign with the static field after stopping the rotational field can be measured and is called T1. When the rotational magnetic field is turned off, the interaction between the spin of the various protons will slightly modify the Larmor frequency of each spin, making them out of phase with the other. This dephasing time is called T2. T1 and T2 give various information about the nature of the tissue present at a specific location [21].

2.2.2 MRI Contrast and Orientation

Various contrasts are available in MRI due to the various characteristic time that we get, as explained before. They have differences mostly based on what can be distinguished. Traditional contrasts are :

- “T1-weighted”: dark CSF / light cord
- “T2-weighted”: light CSF / dark cord / gray matter not visible
- “T2*-weighted”: light CSF / dark cord / gray matter visible

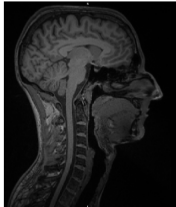
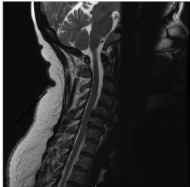
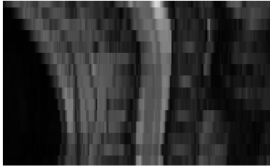
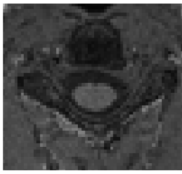
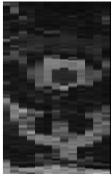
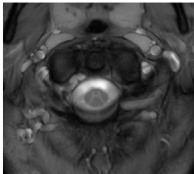
MR contrast	T1	T2	T2star
Orientation, resolution	Isotropic 1x1x1	Anisotropic sagittal 4x0.47x0.47	Anisotropic axial 0.5x0.5x5
Sagittal view			
Axial view			

Figure 2.5 Common MR contrast

As shown here the information contained in the image varies with the available contrast. Indeed T2 will give information about the CSF and the spinal cord. In T1 it is very hard to distinguish these two elements.

The T1-weighted images allow for structural recognition, yet, it is hard to locate the CSF when looking at such images due to the relatively long T1 of the free water. In the T2* images, one can see the white and grey matter within the spinal cord, which is not possible with T1w and T2w MR image. The segmentation of the GM is really relevant in some cases and can help provide more information about the anatomopathology of the condition. The T2* images are therefore critical in the study of some disease such as Multiple Sclerosis (MS) where the myelin content can be affected by the disease. Lesions can be seen through these images [22] and more information about the condition can be known. For example, with MS, T2* imaging allowed for the computation of the lesion load in the WM and the GM which is relevant to know more about the disease stage [23].

However, some new techniques allow for better recognition of the white matter tracts such as diffusion tensor imaging (DTI). This technique is based on anisotropic water diffusion throughout the image acquisition process. This helps to determine the more specific location of fiber [24].

MRI spatial resolution can be anisotropic (high resolution in a 2D plane with thick slices) or isotropic (cubic voxel) and may vary due to slice thickness, the field of view, and matrix size. The matrix depends on (i) the frequency encoding steps and (ii) the phase encoding steps. Both take a single direction of the image plane. The Field of View (FOV) represents the coverage of an MRI image in the various spaces (Fourier, k, and spatial). The typical resolution used in clinical routine is about $1 \times 1 \text{ mm}^2$ in the image plane, and a slice thickness between 1 mm and 3 mm.

2.2.3 Challenges of Spinal MRI

Patients need to remain very still otherwise images can be blurry. Moreover, the CSF around the spinal cord flows back and forth in the inferior, superior direction. This flow can make the spinal cord slightly move as well, which might further blur the image. Unwanted image artifacts might also appear due to geometry distortions and signal attenuation due to different susceptibility in the vertebra or air-filled lung. This artifact might lead to problematic reading and rater bias [25].

The spinal cord also represents a small organ, a typical resolution $1 \times 1 \text{ mm}^2$ is often not enough to grasp the subtlety of spinal cord change, which will be critical in this study.

2.3 Spinal cord lesion and compression: Degenerative cervical Myelopathy

2.3.1 Pathophysiology

DCM is a special type of cervical myelopathy condition which is characterized by the compression of the spinal cord following changes in the vertebrae morphometry. Vertebrae can have a bone outgrowth around its top or base, which is then covered by the closest intervertebral disc. This results in an outgrowth of cartilage pressing on the spinal cord. Other causes can result in DCM, such as ligament calcification. All of these can be seen in figure 2.6 [26].

This results in a decrease in the spinal cord canal size, which will, in turn, cause the compression of the spinal cord itself. This also causes stiffness in the adjacent structure, causing some abnormal activity in the spinal cord. Movement and others will come then because of spinal cord irritation and further compression. This irritation and compression can lead to neural cell loss in the spinal cord and problem along the blood-CSF barrier in more acute cases.

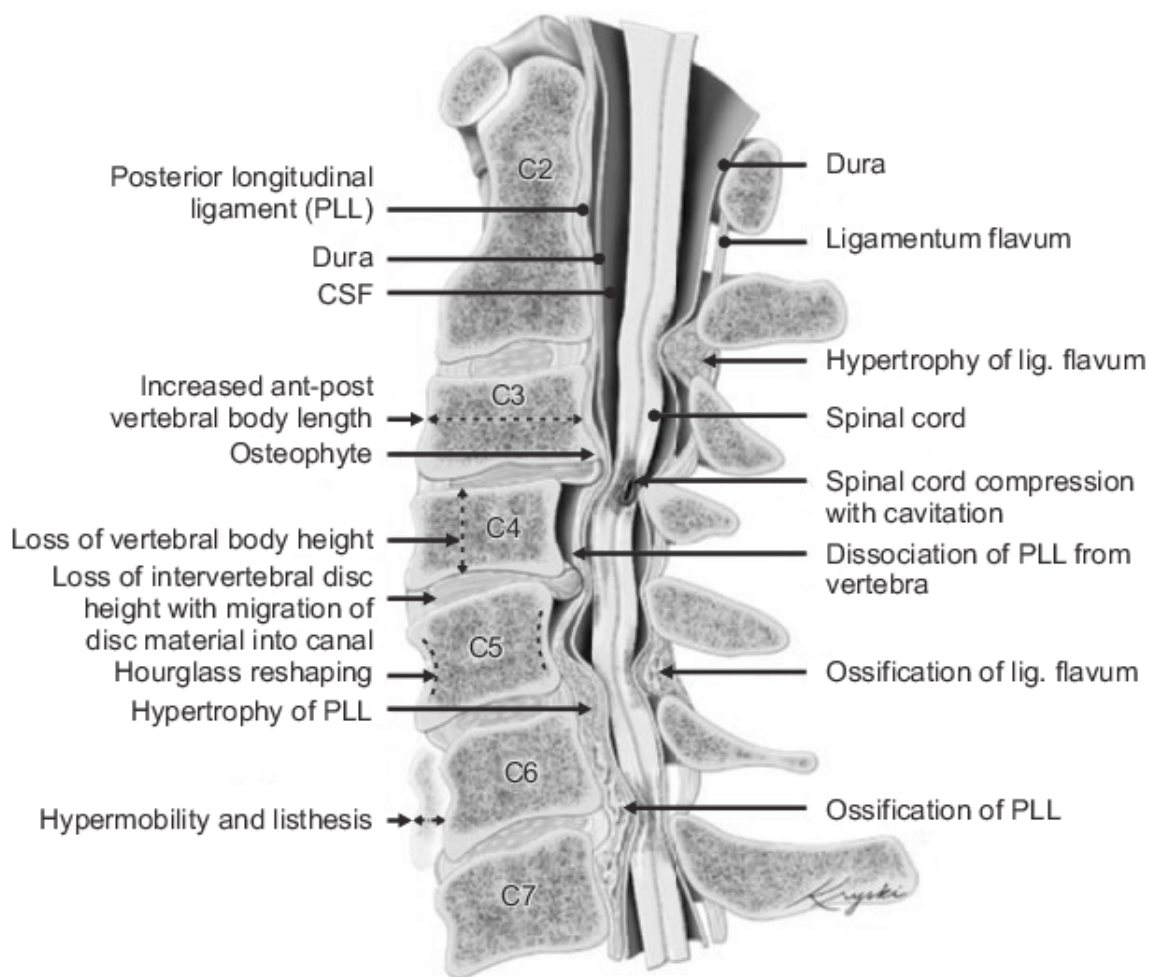


Figure 2.6 Various causes of DCM. CSF = Cerebro spinal fluid. PPL = posterior longitudinal ligament

This represents the various causes of DCM [1]. Since there are multiple causes, there are multiple possible surgeries that we will discuss in 2.3.3.1. This also explains the complexity of the prognosis since the image quality offers only partial information about the cause (for example, bones and disc can be distinguished. Image from [27], license <https://creativecommons.org/licenses/by-nc/4.0/>

The condition is degenerative. This means that the spinal cord compression will get worse and worse as time passes without exterior intervention.

2.3.2 Impact and Diagnosis

The main impact of this illness is the loss of neural cells, which causes motor impairment for the patients. These motor dysfunction often starts with numbness in the extremity and with the loss of distal control further down. One of the issues of DCM is that it is not often diag-

nosed early because of the wide range of conditions that could result in such symptoms [4]. Other symptoms include neck and limb pain, sensory loss, and clumsiness, which comes from the loss of distal motor control. In more advanced cases, the patient suffers from severe motor dysfunction and, in some cases, bowel and bladder control loss. Left untreated, this can lead to tetraplegia.

These symptoms are often mild at the beginning, requiring individual awareness from medical professionals to be diagnosed as DCM in early stages. The diagnosis of the condition takes, on average, two years [26] which is a long time, especially considering the "degenerative" aspect of the disease, which causes worse symptoms over time. For now, a popular method to assess for DCM is the use of a T2 magnetic resonance image, which shows the compression of the spinal canal. It has been determined that a compression of 8–57% can be asymptomatic [11]. The patients are then evaluated of multiple scales such as MJOa and SFI-6D to try and determine their level of motor control.

2.3.3 Current Solution and Issues

2.3.3.1 Decompressive surgery

The current solution is based on decompressive spinal surgery, which can take different forms: Anterior Cervical Discectomy and Fusion: the herniated disc is removed and the vertebrae below and above this disc are fused together using a bone graft.

Anterior Cervical Corpectomy and Fusion: bone growth spurs is removed and the two vertebrae are fused to maintain stability (the affected one and the closest to the growth spurs).

Laminectomy: The lamina is removed, which is the back part of the vertebra that covers the spinal canal.

Laminoplasty: A bone graft is used to enlarge the lamina and so the spinal canal at the point of compression. [28]

The goal of the operation is to relieve the pressure being applied to the spinal cord. The surgery can be posterior (opening from the back of the neck) and/or anterior (opening through the front). There is still an ongoing debate about the anterior and the posterior approach, even though the anterior approach seems to give better results in most cases [10].

The performance of the machine learning analysis tool could be mitigated by this factor, as we try to predict the recovery of the patient without knowing the surgical procedure engaged.

2.3.3.2 Surgery Issues

The risk of such operation is numerous (e.g., dysphagia, pseudo-arthrosis, C5 palsy) [3], but the benefits of successful surgery far outweigh those. The main issue with the surgery lies in its ineffectiveness in some cases. For 25% of operated patients, their condition still worsens after the surgery; for 40%, their condition remains the same, and only 35% of the patients get better [11]. Surgery is considered successful if the MJOa evaluation of a patient improves by 2 points, which is the minimum clinically important critical value for the MJOa scale [8]. Some factors seem to influence the outcome of the surgery, such as age and disease severity. [29]. This does not seem to be sufficient to avoid unnecessary surgery for the time being, which is the default and automatic procedure for cases such as this one, even though risks are great without much benefit in some cases. The current prognosis is based on spinal compression evaluated through MRI and disease severity, measured with the MJOa. However, this only has a limited prognosis value [30] [31].

2.4 Biomedical Data Processing Through Machine Learning

One way of improving this prognosis method could be through the use of newly developed techniques in the artificial intelligence field. Statistics are commonly used in biomedical diagnosis and prognosis. However, the emergence of these new techniques allowed by the constant augmentation of processing power and available data allows for efficient medical machine learning and deep learning model to emerge. These techniques require a huge amount of labeled data to be efficient, as it will be presented here. This section aims at giving an introduction to critical concepts in machine learning as well as an overview of machine learning methods currently available. There are a growing number of studies using machine learning and especially deep learning to process images and data within the biomedical community. These tools can be used for automatic analysis of images (e.g., spinal cord segmentation, lesion segmentation [22]) and data. This helps clinicians and academics with analysis and/or decision-making. These tools allow us to gain precious processing times on repetitive tasks.

2.4.1 Machine Learning Basics

Machine Learning (ML) aims at passing knowledge onto a computer through data and interaction with the user rather than through explicit programming. The goal of such an operation is to provide enough information so that the computer can generalize its definition to an unseen data example.

The field is based on different common concepts. The first ones are the concepts of supervised learning and unsupervised learning. During supervised learning, the machine sees data and gives an output. This output is then compared to pre-established ground truth (e.g., manual segmentation map) with a loss function. The error is then used to help guide the model toward a solution, which is a model that will reduce the output error. Unsupervised learning is based on data similarity and on a pattern that can be found within the dataset. These models are made to group data by likeliness and similarities.

In supervised learning, the model is fed a set of training data. These training data have corresponding ground-truth labels. These will be used by the model to establish its performance on specific data. The resulting error is computed through a loss that is **backpropagated**. This modifies the network to minimize its output error. This will guide it toward an optimal scenario, which will allow for the algorithm to determine the label of unseen data correctly. However, there can be multiple optimal scenarios depending on the chosen loss function [32] that needs to be determined according to the problem. There are two main categories of problem for supervised learning: regression and classification.

The regression models are made to determine a continuous target based on a set of independent features. This represents problems like the severity of a disease.

The classification model aims at predicting the group within a set of categories, to which the input data belongs. For this problem, the data needs to be classified by the user if it is not obvious (e.g., separation between severe and non-severe diseases [33]). Clustering is similar but uses unsupervised learning. The goal of the clustering algorithm is to separate data points in groups based on similarity.

2.4.2 Deep Learning

Deep learning regroups specific methods from machine learning. These methods rely on Artificial Neural Network (ANN). ANN is a programmed system made of multiple neurons, which are functions (filter function such as convolutions and activation) that apply some logic to the input to give an output. These neurons are arranged in layers that are connected. An ANN has an input layer, where the data is fed, followed by a single or multiple hidden layers and then an output layer where the prediction is formed. The connection between the neurons is influenced by weights. If a neuron provides a lot of information, it should have a heavier weight than the one providing less information. The weights and exact parameters of each function are determined through the training. The model is then validated and tested on unseen data. Deep Neural Networks (DNN) are ANN with more than five layers. These

architectures are often used for more complex problems as each layer provides a new set of parameters to be modulated. This method usually allows for less manual features extraction as recent networks are made to detect patterns and retrieve useful information from images and data. For supervised learning, during training, the model modifies its weight based on the data fed and the expected output.

Training is an iterated two-step process: the first one is the forward pass, during which data is fed to the network that gives an output. The loss value is then computed based on the established loss function, the output of the network, and the ground truth, which represents the expected output. The loss is then used to perform the backpropagation of the error, which will influence the way the neural net changes its decision process. Since the model is generalizing from examples, the quality and coherence of the data are of paramount importance. The more cases the network sees, the better it will perform on similar cases it has never seen. Validation is a way to evaluate the performance of the model on unseen data after each training cycle. However, to avoid adjusting user-defined parameters during validation, another set of data called a "testing set" is used to measure generalization performance.

2.4.2.1 CNN

Convolutional Neural Networks (CNN) are special neural networks that use a specific set of functions in their neurons and, therefore, specific layers.

Convolution layer: This layer uses convolution to detect edges and shapes within a bigger image. The kernel size of the image determines the size of the receptive field of the filter. The dot product between the filter and the receptive field on the model gives a cumulative score that represents the likeliness between the filter and the receptive field. The score of each receptive field is stored in an activation map. This map will represent where each shape is located in the image. The filters are often randomly initialized as the backpropagation will guide its shape later during training.

Pooling Layer: The pooling layer is there to reduce the size of the activation map. This helps to avoid overfitting by suppressing some little details to keep the big picture clear. There are different types of pooling such as max-pooling (where we keep the highest value to represent multiple receptive fields) and average (where we take the average to describe multiple receptive fields).

Fully Connected Layer: These layers, also called dense layers, are a collection of neuron that acts like binary gates. Their influence is created through the weights they hold in their connection to the previous layer. They will output some features (0 or 1) based on the existing shape and pattern in the input. The output of such a layer is the probability of the presence of each shape or even a category in the case of classification. Their hyperparameters are the number of neurons which will change the number of the different element the network could perceive and use for classification.

2.4.3 Application in Biomedical Imaging

There is a wide range of various machine learning and deep learning techniques available nowadays for medical images and data processing purposes.

These methods are efficient and automatic and often stem in a coding competition. The growing interest in machine learning, as well as the performance these methods offer, explains the explosion of the number of studies aimed at developing and applying new and efficient models to the biomedical field. The application is numerous and is developed to be quite complex along the years, as shown in 2.4.3.1, 2.4.3.2 and 2.4.3.3. Among these methods, deep learning and the help it provides to process natural images, have grown to be very common in the biomedical community. As mentioned in 2.2, the MRI imaging technique is useful to detect a lot of anomalies with soft tissue and more within a patient's body. MRI images are, therefore, a good field of application for the existing deep learning techniques [34], as the problems are numerous, and data is collected for diagnosis. Even though some deep learning methods exist to help acquire images, this section will be focused on more downstream applications and overall MR image analysis.

2.4.3.1 Segmentation

One of the main applications of deep learning in biomedical imaging is segmentation. The goal of the segmentation model is to classify every pixel on the input image as part of the background or part of a specific class. The model outputs a segmentation map, which is an image that can overlay over the input to give you information about the shape, presence, and localization of some targeted element [35]. This allows for quantitative analysis over the image. Most DL segmentation methods rely on CNN, such as brain tumor [36] or pancreas segmentation [37]. The most common model is the Unet, which is a CNN using an encoder and a decoder as well as a skip connection. Those skipped connections are often used in the biomedical field [38]. Unet and its variation are still widely used in competition as its

performance is really good when it is well trained even though it was created in 2015 [35]. It has spawned a lot of variations that are often more application-specific. Keypoint detection is close to segmentation as its goal is to determine the position of a structure within the image.

2.4.3.2 Diagnosis

Deep learning models are also developed to be able to diagnose a specific disease based on images and clinical data. This tool is made to help clinicians and is starting to gain some trust as its performance becomes unmatched [39]. One way of improving diagnosis is to improve screening detection as it was done for breast cancer, which is a big challenge. It consists of creating a neural network that can learn to detect tumors to classify mammograms from patients [40]. This is possible as tumors represent a recurrent pattern in cancer patients. Recently one system developed by Google seems to have even improved on what human doctors can do. Performances are crucial for the model to be trusted; the main advantage that has an efficient system over a human is a lack of bias and fatigue when screening this type of image.

Closer to neuroscience and our problem is the diagnosis of multiple lumbar pathologies by a single Fully Convolutional Network (FCN) [41]. The network will detect various key points, such as vertebrae and intervertebral discs. It will then analyze regions of interest and use the information to conclude on the state of each patient. This is a tedious and time-consuming process for the doctor, which can be improved and made less complicated using deep learning.

2.4.3.3 Prognosis

Prognosis models are created to determine the future state of a patient based on its current state. It can be crucial for practitioners in the decision-making process, especially in the case of risky procedures. The 5-year survivability is a standard metric for dire diseases such as cancer. As cancer data are numerous, it is easier to create models based on these data. Multiple studies aimed at predicting the 5-year survivability rate of patients through ANN or more common ML techniques such a Support Vector Machine (SVM), which is a learning algorithm that finds one hyperplane to divide data into two categories [42].

Methods using neural networks are also more and more common for patient prognosis and have unmatched performances. Cancer is again a great example for such method, as new deep learning approaches provide good results in multiple cases [43] and even exceed any previous method in some cases (e.g., breast cancer [44]). Deep learning model allows for the use of multiple inputs with different data types, which can be useful and provide very

interesting results in different cases [45].

This method can help make decisions but also discover the correlation between multiple independent variables and might help improve treatment for future patients. There are already treatment optimization through ML and thanks to ML diagnosis [46].

2.4.4 Machine Learning Processing of DCM

Machine learning and deep learning are already great tools for prognosis and diagnosis in some cases, such as cancer. It launched multiple studies that try to perform prognosis on different illnesses where it will hold significant clinical value. One of them is DCM, which has now been the focal point of four machine learning-based studies over the past three years. One of these studies aims at diagnosing DCM in patients with MR imaging data. The analysis is done by a deep learning method with a neural network trained on healthy control and patient. For this study, they also prepare a second neural net to predict the MJOa score of each patient. However, this is a pilot study that was only trained and tested on 28 patients, which is a meager number for these learning methods, yet it gives encouraging results. Their diagnosis model was able to give a 91.7 % accuracy while the Prediction of MJOa score only differs by 0.714 from the established ones. [47]

There are two studies focused on processing AOspine clinical data that contains a certain amount of information on patients such as age, sex, weight, MJOa pre-surgery, duration of surgery. Both studies aim at performing post-surgery prognosis, which is the main goal of this study. Both studies used different models but used the same data: the AOspine Dataset. Merali et al. tested various machine learning model and an artificial neural net. Best performances were obtained from a random forest model, with a classification score of 78% [5]. Khan et al. got similar performances with a generalized boosted model as its best performer. [6] Both studies considered, among their criteria, surgery information such as operation length and level, which can be unknown pre-surgery criteria for real-life applications. [5]

The last study used different data as they based their model on features extracted from DTI images. This study also aimed at predicting post-surgical outcomes. They did not aim at predicting the class based on straight MJOa difference but preferably on what they called recovery ratio. Which is calculated as follows :

$$\text{recovery ratio} = \frac{\text{postoperative score} - \text{preoperative score}}{17 - \text{preoperative score}} \times 100\% \quad (1)$$

As mentioned in 2.2.2, DTI metrics can offer a more detailed view of the tracts and parts inside the spinal cord. Authors determined nine regions of interest, which are two in the gray matter (GM) lateral column (LC) ventral column (VC) and dorsal column (DC). The DTI metrics were then obtained for each of these regions. The most accurate model was a least square-SVM, which is an SVM where the distance to the separation line is calculated using the least-square formula. They obtained 88% classification accuracy. However, they were using an entirely different and private set of data [48].

2.5 Contributions

During this thesis, I took part in the development of multiple tools within the NeuroPoly laboratory that are now available. The laboratory is developing a toolbox (Spinal Cord Toolbox) for the analysis of spinal cord MR images. This toolbox was of huge help for the preprocessing of this study's data as it is for many people in the community (more than 150 citations). However, some functions and functionality could benefit from updates and improvements. These contributions are added to this study. All of these are detailed in this chapter.

2.6 Spinal Cord Toolbox

Help was provided for the implementation of new functions that were useful for this work, such as completion of current disc labeling. This improved the existing function by allowing correction of the existing label instead of requiring manual labeling of the whole image. It let the user open a label image on top of an anatomical image to modify the name by simply clicking on the key points that are needed. This function was useful to create ground truth labels for the second contribution.

The spinal cord toolbox also offers the possibility to straighten the spinal cord, which returns a topologically similar image with a straight spinal cord [49]. However, this operation changed the value of the label file and created artifacts. An option for avoiding these issues that use nearest-neighbor interpolation method on dilated labels was implemented.

Windows OS users were assisted by creating relevant documentation for the use of the toolbox within a docker container and later inside the Windows subsystem for Linux [50]. I maintain the repository for the windows user, which required minor updates for the correct

installation of a new release of the software.

2.7 Vertebral Labeling by Deep Learning

Spinal cord toolbox offers a function to label vertebrae using template matching [51] on 3D data and detection of C2 vertebrae using a HOG-SVM (histogram of gradient-SVM) model. This method, however, failed about 10 % of the time due to some missing point or missing detection of C2, which causes the entire program to fail. The use of deep learning to detect intervertebral discs is more and more common. Therefore it was only logical to apply new methods to this problem. This entails implementing, optimizing, and testing a neural network in Pytorch, which is one of the main frameworks to develop DL applications. Different models were tested and evaluated, such as inception modules, Countception models, and attention U-net [52] [37].

This study revealed that the developed model outperforms the existing solution using a modified Countception model [52], which is a fully convolutional network (composed only of convolutional layers) that was first used to count cells in microscopic observation. It is described in figure 2.7.

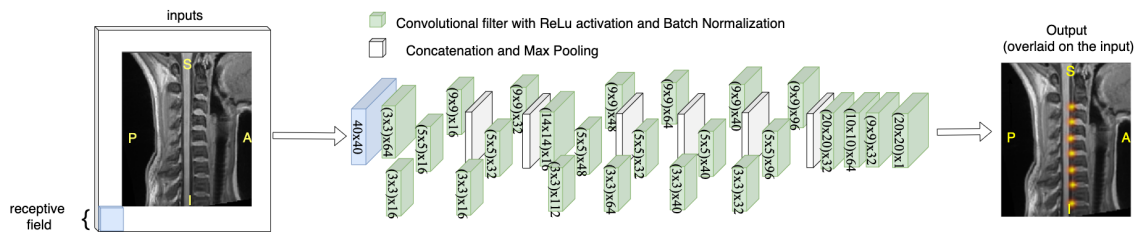


Figure 2.7 Model with its receptive fields (40x40) on the input. [12]

This model was trained to use the spine generic public dataset [53] and then a private dataset that needed to be manually labeled to create examples for the network to train one. This was the reason for the contribution in SCT described previously. In total, the network was trained on 435 T2 and T1 MR images. To improve it, template matching on the heatmap output by the model was then introduced.

2.8 Prognosis

This thesis offers a new approach to prognosis. As explained in subsection 2.4.4 some studies already exist for the post-surgery of prognosis of the DCM. However, these studies mostly present clinical data analysis and often use operational data (e.g., duration of surgery and operation level) to establish their prognosis. The main focus was on creating a model to perform prognosis using only relevant pre-surgical features to which we add multi-modality image analysis and feature extraction, which was not reported before. The impact of the use of surgical data in the analysis was also observed. The impact of multimodal MR image analysis added to the machine learning and deep learning model were studied. One of the options to exploit MR images is to retrieve features from the images that can be utilized in a clinical setting to improve on the current prognosis methods.

CHAPTER 3 MATERIAL AND METHOD

3.1 Material: Dataset presentation

This section provides a description of the data that are used in this study. Machine learning and, by extension, deep learning are example-based learning models, which explains the critical importance of the dataset constitution as it will drive the learning process.

3.1.1 Dataset Constitution and Specific

This dataset was prospectively acquired as part of the AOspine CSM North America, [54] and AOspine International CSM study [55] led by Dr. Michael Fehlings from the University of Toronto. This study aimed at providing a multi-center dataset for the analysis of DCM. It regroups 769 patients of which 485 completed the three follow-up after the surgery and had filled the features that were used. They were included in the study considering various inclusion and exclusion criteria:

1. The patient is more than 18 years old.
2. The patient suffers from **symptomatic** DCM with signs of myelopathy, which was evaluated through his motor impairment.
3. Patients did not undergo previous surgery for similar condition.
4. The patient is willing to participate in the study and its follow-up and able to understand the local language.
5. The patient does not suffer from problematic comorbidities (e.g., active infection, neoplastic disease, rheumatoid arthritis, ankylosing spondylitis, concomitant symptomatic lumbar stenosis, trauma).
6. The patient does not suffer from a condition that would hinder accurate evaluation (eg., neuromuscular disease, significant psychiatric disease).

All accepted patients underwent surgery at the center where they were recruited. Before the surgery, each patient filled out a form with an investigator detailing their clinical and personal situation (e.g., comorbidities, smoker status, marital status). The investigator also evaluated the patient's condition on the clinical MJOa scale, which represents the

patient’s motor impairment between 0 and 18. Other grading systems were also used (e.g., sf6d, NDI.). The first set of data also contains information about patients’ surgery, such as operating procedures (anterior/posterior approach, operated levels), and observed/reported complications.

After the surgery, there were three follow-ups with the patient at 6, 12, and 24 months. During these interviews, the investigator reported the new clinical scores of the patients and evolution or apparition of complications.

All these data were compiled on a table. Information about the availability of MR images for each patient was added, representing 266 data points for each patient.

The MR images were mostly sagittal, and axial acquisition focused on the cervical parts. The availability of MR images (sagittal and axial), their field of view, and contrast were not consistent. There were no agreed-upon acquisition protocol which explains the high variability in the image quality and availability. However, this gives a good heterogeneity to our data creating a dataset close to real life condition.



Figure 3.1 Example MR images from the studied dataset

This shows different MR images examples from the AOspine dataset. Green arrows points to the compression area in each patient. As shown here, it is possible to witness multiple compression on a single patient and even on the same level. You can also grasp the heterogeneity of the MR images.

3.1.2 Patients and Feature Selection

During the study, we aimed at verifying the added value of MR images to the clinical data of the patients. Moreover, some of the collected data were not relevant to our problem. Several transformations to obtain a smaller dataset were performed. The resulting set contains pieces of information more relevant to the patient’s health condition (comorbidities, age, duration

of symptoms). It depends less on arbitrary criteria (date of surgery, date of waiver). Too much data would allow the model to learn the specific of each patient, which would cause the model to overfit on the training data. Multiple datasets representing part of the findings gathered for the AOspine study were constructed. They are detailed in the next section. The goal was to avoid removing too many patients as our dataset is already relatively small for deep learning and machine learning analysis.

3.1.2.1 Patients Removed

This work presents different approaches that were tested. These approaches require different types of data. The first approach aims at exploiting only clinical data, the second one focus on deep learning to obtain information from both T1 and T2 images, and the third one focus on feature extraction. These three approaches require different data (e.g., our second approach is based on the exploitation of T1 and T2 images), which explains the extraction of three datasets from the available raw data.

The first dataset regroups the clinical data of the 485 patients that completed the study. These data are used for machine learning processing. Since there is no constraint on image availability and resolution, there are more patients to consider. This dataset will be used with the first approach.

Criteria based on the analysis of the MR images were added. A second dataset composed of 260 patients for which we had T2 and T1 sagittal images was therefore created. This dataset was exploited to test multiple networks using the MR images and clinical data. All images provided a resolution of at least $1 \times 1 \text{mm}^2$ in the sagittal plane, which is important to avoid losing information in the preprocessing phase that is discussed in the section 3.3.

The last one regrouped patients with high-resolution axial MR images (resolution better than $1 \times 1 \text{mm}^2$ in the axial plane) that covered the compression region. This was done to extract more precise features from the MRI that will be exploited in machine learning analysis for our third approach. This dataset regrouped 158 patients.

This dataset division is summarized in the following figure.

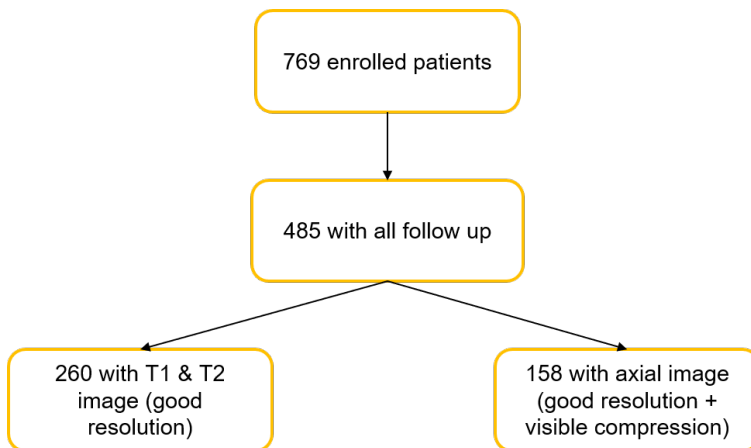


Figure 3.2 Dataset division

Representation of the dataset creation. An image is considered having a good resolution if its resolution is above $1 \times 1 \text{ mm}^2$ in the considered plane.

3.1.2.2 Features

Features allow for separation in two different datasets: one that contains pieces of information about the operation and one that does not. For the dataset with the operation information, we decided to keep the reference of the disc affected by the surgery, the duration of the operation, and all information about comorbidities as well as original clinical scores. The goal was to see if surgical information influence classifier performance. In all dataset, information about the selected approach and type of operation were kept as this is decided before the operation by the surgeon. However there is no guarantee that it was the best surgery to perform. All types of surgeries and surgical approaches are represented. The two approaches are well represented. The type of operation is a bit more complex as these are not mutually exclusive and might concern more than one spinal body.

3.1.3 Dataset Statistics

These statistics are computed on the clinical data for the 485 patients that completed the study. To provide a relevant real-life scenario, it is needed to have a balanced group of subjects for such a study. Some of the statistics that described our cohort are here detailed. Among the patients, 62.8% are men, and 37.2% are women. Five hundred and fifty-four are smokers. This criterion seems to be correlated with less successful surgery [29]. This appears in our dataset with only mild importance as the correlation is close to 10% between these criteria and successful surgery.

A wide range of age is represented. This is important as this condition affects mostly adults that are more than forty years old, and the probability of the condition increases with age.

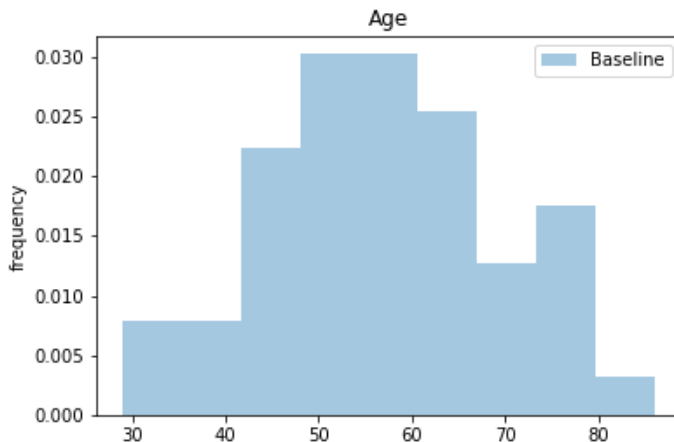


Figure 3.3 Age distribution

As shown here the age of the patients ranges from 30 to 84 which represents a good range for such study as age is an important factor in the patient's recovery.

We can see on the figure above that there is an important number of patients that are roughly 75 years old. The low number of data from patients that are 80 years old or older could be explained by death from other diseases or natural causes before the development of this condition. Surgery still remain relevant, as the condition is degenerative. Even if the surgery is not really successful, this could still halt the disease progression and allow the patient to keep some of his sensorimotor skills. This is especially relevant as the life expectancy continue to increase, as could be used for future research.

Age is a very relevant features to our problem, but this is not the only one. There are indeed some correlation between some of the features that are included in the dataset and the surgery outcome as seen in previous study [29]. These features are included in the exploited clinical data for all three approaches. They indeed seem to be important for the post-surgery prognosis and are based on objective information, whereas patient evaluation can be subjective as we will discuss in the next paragraph.

3.1.4 Patient’s Evaluation and Outcome

The MJOa scale used in the AOspine study evaluate the patient’s skill through physical tests and information provided by the patient on four different categories: motor dysfunction of the upper extremities, motor dysfunction of the lower extremities, sensory loss, and sphincter dysfunction. The MJOa score depends on the patients state the day of the evaluation and on the investigator (who can change from patient to patient) which means that it can be biased. This is, however, the case for the evaluation of each patient , as they are not done over a long period of time.

The surgery can either be successful or unsuccessful. in this study, surgery is considered successful if the difference between the preoperative MJOa score and the postoperative MJOa score after 24 months is greater or equal to 2. This is determined with the minimum clinically significant difference [8]. However, the same investigator performs the pre-surgery and post-surgery evaluation, which means that the difference between the two scores is less biased. Moreover, it is widely known in the community that a difference of 2 represents a significant improvement of the patient’s skill. The MJOa difference will be our ground truth for this study. We will classify it following this criterion $\Delta M_{joa} \geq 2$. This allows us to create two functional classes.

Roughly 63% of patients have had a successful surgery following our criterion of ($\Delta M_{joa} \geq 2$) in the first two datasets, and are therefore, positive samples. The third dataset contains only 57% of positive cases.

3.2 Approach of the Problem

Based on the data and on the literature review, it was decided to perform a three-prong approach: the first one is to adapt ourselves to what already exists whilst studying the influence of surgical parameters. These elements were taken into account in the previous study. This is on par with Merali et al. and Khan et al. studies; however, here we considered pre-surgical data only [6] [5].

The second part consists of trying to exploit the MR images. Firstly, with the images as sole input for a deep neural network to train on. Both modalities were used as a training example. The single T2 image, which is more readable as the CSF is fairly visible, is used as input for an out of the box Resnet model. This was done after different preprocessing steps such as spinal cord straightening and cropping, which will be detailed in the section3.3. Multiple models were exploited based on the state-of-the-art image classification models such as Resnet [56], and a custom Multi-Input CNN based on inception network [57]. In a second

time, the study moved toward the joint exploitation of clinical data and MR images, to try and improve on the previous research by adding information from the clinical pre-surgical data. The clinical data was processed by a neural network.

The third part is based on semi-automatic feature extraction such as weighted average in the various white matter tracts. This gives information about the associated image and extracted features are added to the clinical data to evaluate the new features relevance and the model's performance. This aimed at getting closer to the data used in the DTI study, which, as the literature shows, exhibits better performance on their dataset.

3.3 Image Preprocessing

Image preprocessing is a common step for deep learning study. In a multi-center dataset, there is often a high inter-patient variability due to a difference in available machine and acquisition parameters. As the deep learning model tries to recognize patterns, the idea is to avoid any non-illness-related pattern in the image. Pre-processing steps are often semi-automatic, and create the final input that will be fed to the network. Various preprocessing pipelines were used to improve the performance of the model. Pre-processing pipeline can also be used to extract data and features from the images.

3.3.1 Spinal Cord Segmentation

The spinal cord segmentation is useful to get information about the condition, such as the compression percentage, that can be obtained by looking at the cross-sectional area of the spinal cord. It can also be useful for other automatic methods such as spinal cord straightening, which is detailed in 3.3.2 as well as template registration. In any case, it is providing information about the inferior-Superior FOV and can be processed to obtain the Cross-Sectional Area (CSA). This area represents the size of the spinal cord on any axial slices. This was mostly done using the Spinal Cord Toolbox (SCT) [9] and specifically the deepseg method that is based on deep learning segmentation of the spinal cord in the axial plane with a Unet [22]. Obtained images were then manually corrected if required.

3.3.2 Spinal Cord Straightening

Spinal cord straightening is used for the template-based analysis of spinal MR images. In SCT this is done using an iterative B-spline method [49]. This straightening provides an image that is normalized in voxel value and provides a reduced FOV centered on the spinal cord, which will help focus the attention of the neural net on the compression. This could

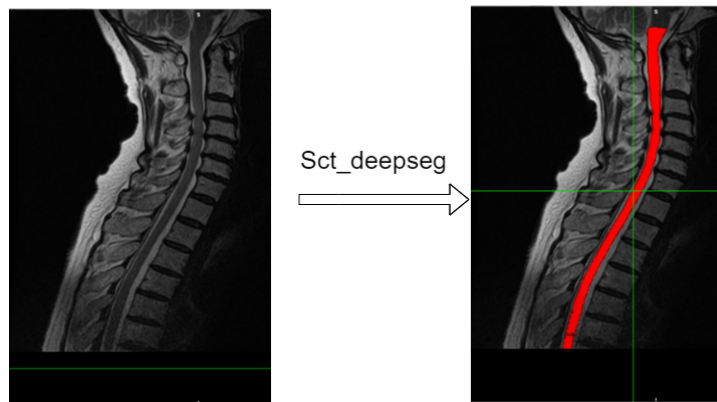


Figure 3.4 Application of the `sct_deepseg` method on a T2 sagittal image from the AOSpine dataset

help avoid overfitting by removing the non-related pattern.

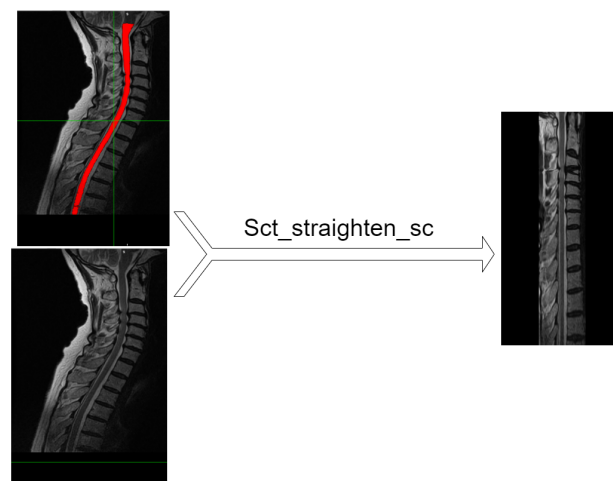


Figure 3.5 Straightening example

Sagittal T2 (left below) with its segmentation (top left) that gives the straighten image that is on the right.

3.3.3 Image Cropping

Image cropping was performed on both straighten and original images. This idea is to reduce the IS FOV around the main compression point. This was done manually using an SCT function. All images were resampled to 0.5 mm isotropic, to increase resolution and unify physical distance, and then cropped to 150 voxels along the IS axis. It is important to keep images size constant, as the input fed to the network for training requires a fixed size

batch of multiple images.

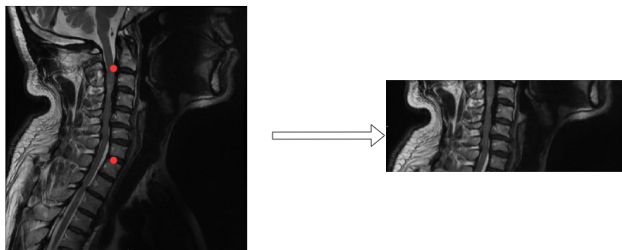


Figure 3.6 Cropping example

Sagittal T2 (left) with the point selected for cropping (red) and the results of the operation (right).

3.3.4 Inter-vertebral Disc Labeling

Inter-vertebral disc labeling is useful for registration function. Moreover, since the DCM starts with disc outgrowths, these represent a region of interest to our study. This was done using a newly implemented technique. The method relies on the first detection of the C2/C3 disc by the HOG-SVM technique, then the application of straightening followed by the localization of all disc by deep learning, as explained in 2.7. The template matching avoided false negative while improving results precision. The exact position of the label, at the posterior tip of each disc, was done using the heatmap.

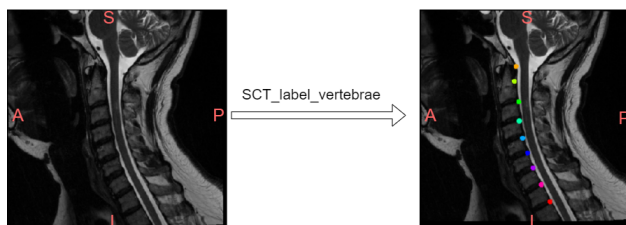


Figure 3.7 labeling example

An example of intervertebral disc labeling. Each dot represents the detected position of a disc. Its value gives the number of the disc according to our convention. The convention is that the disc takes the value of the next vertebrae (e.g., the disc between C2/C3 takes the value 3). Labels are represented by a single pixel, but in this example they have been dilated to improve visual quality.

3.3.5 Feature Engineering

Features engineering is part of most machine learning studies and aims at creating new information based on available data. In this study's the creation of new features is mainly

based on the analysis of MRI images, which yield tabular data that can be added to the clinical one already available.

3.3.5.1 Compression Area

The Cross-Sectional Area (CSA) is obtained by measuring the segmentation of the spinal cord in the axial plane. This can be accomplished by level, if the disc is labeled, or by slices. SCT can compute the CSA for a subject based on these elements.

The CSA can give multiple information that was a note given otherwise. This allowed us to get closer to the current prognosis method by retrieving the compression percentage. This was retrieved using the minimal area, which represents compression, and the mean area of the spinal cord. If the level is labeled, it can also give the region of compression, which can help determine the region affected by the surgery.

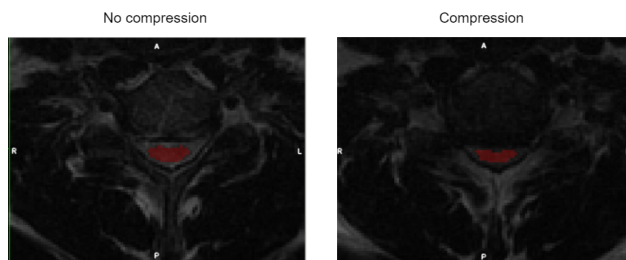


Figure 3.8 Compression example

The cross-sectional area of the spinal cord decreases in the compression region

3.3.5.2 Atlas Information

The use of an atlas could be of some help to get information about the various tracts that run through the spinal cord, as presented in 2.2. The PAM50 [58] template is used for this study. It contains 43 different 'labels' that are specific tract segmentation. The registration algorithm used is 'column-wise,' which perform a "R-L scaling followed by A-P columnwise alignment" (from SCT's wiki [59]). This is recommended for the compressed spinal cord as it allows for the deformation of the segmentation. This relies on the previous segmentation and uses the position of each disc.

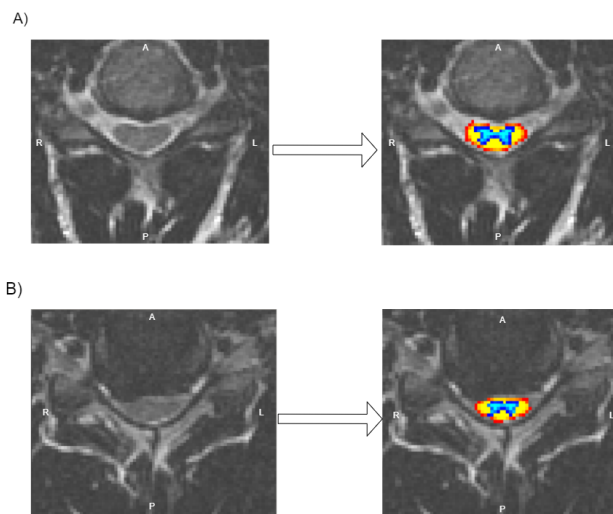


Figure 3.9 Example of the atlas registration

On both image, the White matter (yellow/red) and the gray matter (blue/light blue) are represented. These are issued from the PAM50 template registered to the image. The image A) shows a part of the spinal cord without compression, while B) shows a compressed area. Both images are extracted from the same subjects. A, P, R, L are respectively Anterior, Posterior, Right and Left.

After that, the weighted average of each pathway in each slice or each level is extracted.

3.4 Model and Optimization

The different Machine Learning algorithms tested will first be detailed in this section, followed by a description of the Deep learning models that were used in this study. The integration of clinical data in these models will then be explained. DL models are probabilistic, which means that they could provide additional information on the certainty of the prognosis. The optimization process will be detailed in the next subsection.

3.4.1 Machine Learning

Machine learning models are various and already implemented in the Scikit-learn package [60]. They consist of multiple techniques, such as decision trees and gradient boosting. The classifiers are deterministic, which means they don't give information about the certainty of the model of their answers.

3.4.1.1 Random Forest & Extra Trees

Random forest (RF) is a machine learning algorithm from the “Bagging” technique. Bagging is based on multiple objects (here decision trees) each giving an output from an input. Every one of them then votes for one category or the other (in the case of binary classification).

RF is based on the use of multiple binary decision trees. Each tree gets assigned a random subsample of parameters from the input. It then returns its answer, which is aggregated with all the others. The majority decides on the category. It differs from the "bagging classifier" because not all features are considered on each decision node [61].

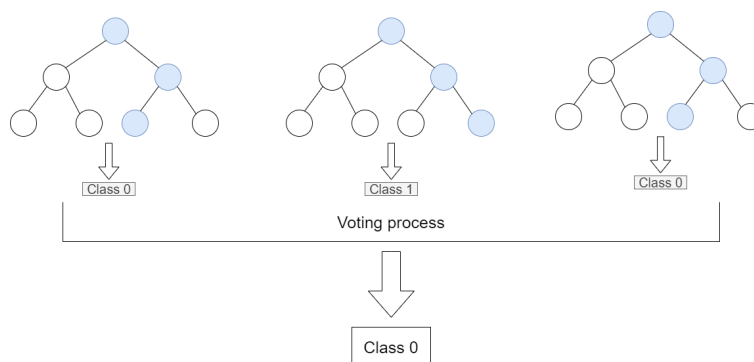


Figure 3.10 Random forest visualization with three trees

Each tree receives a subsample from the dataset

Extra trees also used multiple binary decision trees. In this model, all trees get all the input but the depth of each decision tree is randomly selected. This model offers a different point of view to a similar strategy. Random forest was the best performing algorithm reported by Merali et al. [5] which explains why it was tested in our study.

3.4.1.2 XGBoost

XGBoost is a 'Gradient boosting' algorithm. Boosting technique is based on a decision tree as well, but differs a lot from the previous algorithms. They use the past results, as a loss in deep learning, to select the features that are used as input for the tree. The gradient boosting technique uses the gradient descend algorithms to minimize the error [62]. Its performances have already been proven on biomedical problems [63].

3.4.1.3 K-Nearest Neighbors

This model is based on the creation of clusters, which are groups of points, with the training data. Here there are two possible outcomes; therefore there can be two clusters. This could

be represented by an n-dimension plane (n features). When trying to predict from unseen data, the point is placed on the “graph.” The algorithm then computes the distance between the new data and the training points. The prediction is made based on the “K” nearest neighbors. The predicted class is that of the majority.

3.4.2 Regularization & Overfitting

Overfitting is a well-known paradigm in machine learning and deep learning. It is based on the fact that the model memorizes example data instead of generalizing the pattern within them. It is noticeable by perfect accuracy in the training data and poor accuracy of the validation and testing data. Overfitting can also be manually reinforced by optimizing model hyperparameter to improve validation accuracy, without having any performance improvements on the testing set. Regularization is a common method in machine learning to avoid this problem. It aims at reducing complexity during training by modifying the input image or the weights of the neural net. Here we used dropouts [64], Global average pooling, batch normalization [65], and regularization terms [66], which help the model generalize.

Data augmentation can also help with overfitting, by artificially increasing the size of the dataset. For our approach we used flip and rotation transformation to create “new” images that were not seen before by the network. We don’t want to use any transformation that would modify the original image too much as the compression needs to remain fairly visible.

3.4.3 Deep Learning Architecture

3.4.3.1 Resnet

The first network tested was Resnet. [56]. This architecture is state of the art in the term of image classification as it is the model with the best performance on the ImageNet classification benchmark. It is a deep learning model with multiple configuration existing ranging from 32 to 1001 layers. Resnet most used configuration is the 152 layers (ResNet 152). It uses skip connexion which helps improve classification results as well as the speed of the training process and is important in medical image processing [38].

This model is quick and efficient; it seems a good starting point, especially since it is involved in other biomedical projects such as skin lesion classification [67]. This model is made for single image classification and was used as such with the T2 MR image as input. The actual implementation was validated through a test on the CIFAR10 dataset, which yielded expected results before Hyper-parameter optimization.

The number of ResBlock can be modified, which changes the number of layers to create

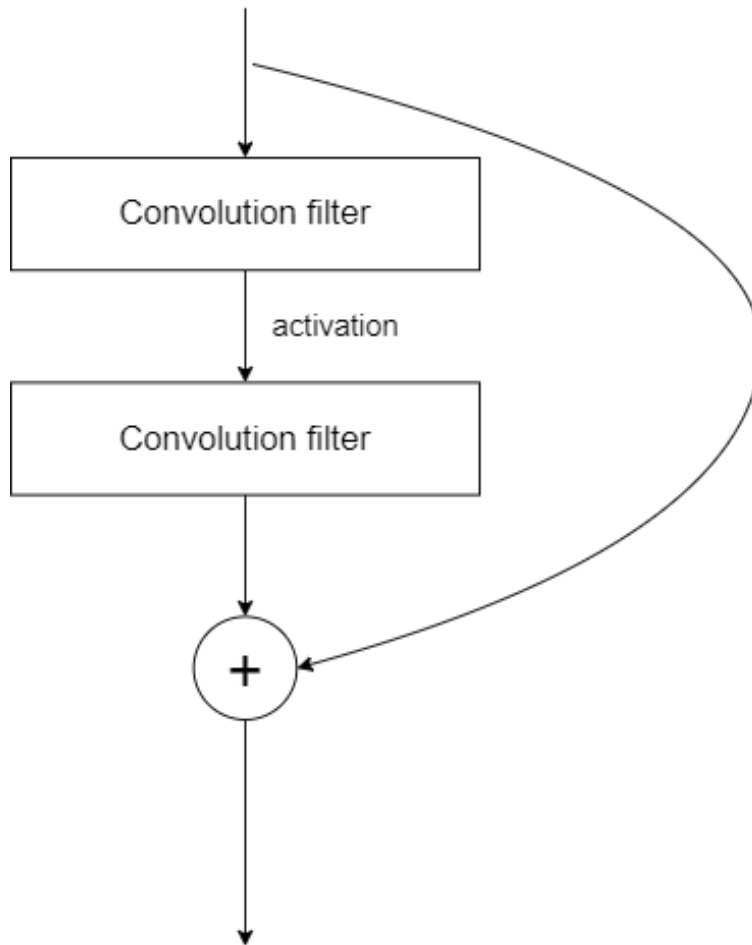


Figure 3.11 Example of skip connexion in a deep neural network

deeper architectures. Kernel size and strides, which are parameters of the convolution layers, were also modified during the optimization process. Kernel size represents the size of the filter applied to the image, while strides represent the way the filter is convoluted with the image, representing the number of removed pixels due to the filter application.

3.4.3.2 Multi-Input CNN

The second architecture was a custom-made Multi-input CNN. It was built with three branches that transform the input before the classification step. Two of these branches were similar as they were made to obtain features from the T1 and the T2 images. The association of T1 and T2 images might give information about the surgery outcome. It has been shown that T2 hyperintensity associated with T1 hypo-intensity in the compression area often results in poor surgical outcomes. [68]. The model consists of multiple 2D convo-

lutional filters followed by Rectified Linear Unit (ReLU) activation layers. Layers offer various hyperparameters, such as kernel size and strides. Strides will divide the image in width and heights. The third branch consists of dense layers and batch normalization applied to the clinical data. We added Dropout Layers to help with regularization. These layers drop features with certain probabilities helping the network avoid overfitting and producing more robust results. The chosen branches are changeable to research the way to automatically retrieves essential features from the images. The combination of T1w, T2w, and clinical data features was done through a concatenation after the encoding. The model outputs a 2-class prediction created by a softmax activation. The current network is not made to be used with a missing modality.

3.4.3.3 Inception Module and Network

The inception module consists of different filters and processing applied to the same input image, which are then combined. This improves processing by offering to the network different ways of processing the image that is then concatenated. The various paths offer different filter sizes which allow the network to focus on a smaller part of the image. This was first introduced in the inception-v3 architecture from Google for image classification. [57]

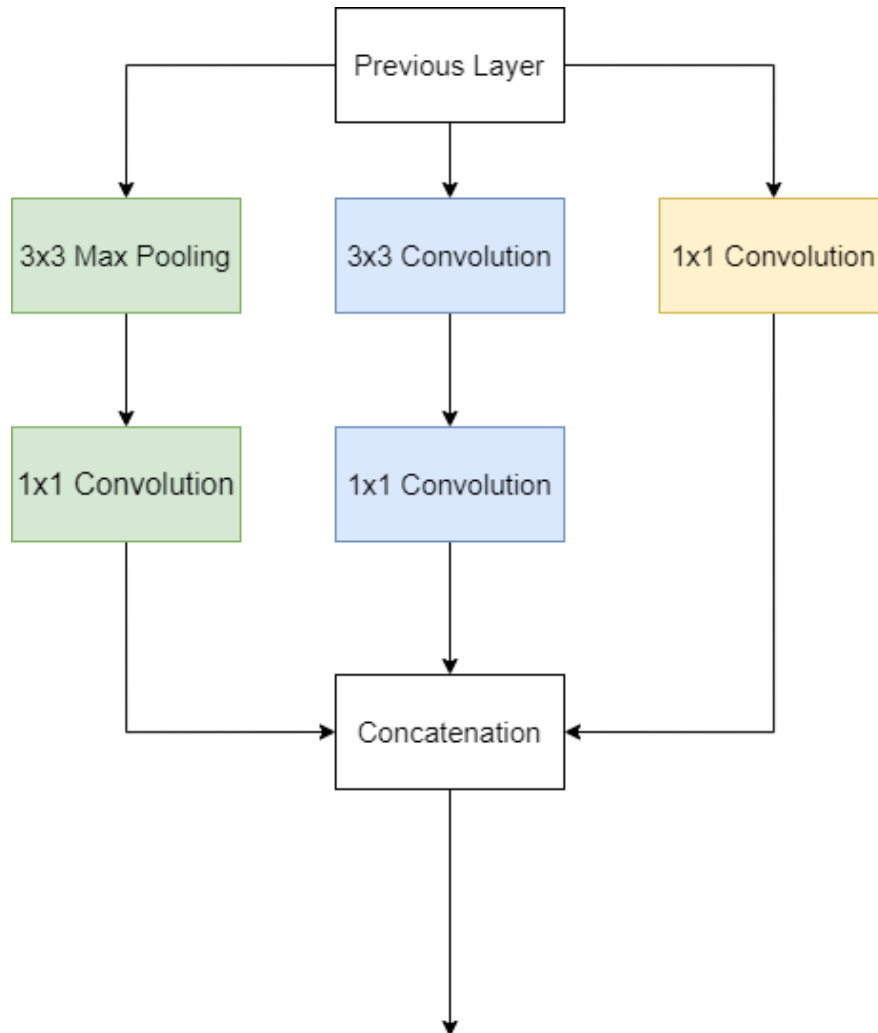


Figure 3.12 inception module

On this module you can see the three ways available that will retrieve different features from the image. The model will select what part is the most relevant to pass it further down.

3.4.4 Training and Model Selection

3.4.5 Machine Learning Model Selection

Machine learning model selection was made with a ten-fold cross-validation evaluation of the different models which will be detailed here. Cross-validation is commonly used to validate a model when the quantity of data is low. The data is usually divided into training, validation, and testing. K-Fold cross-validation relies on permutations between training and testing. We split the data into K groups called folds. We then perform several training on K-1 folds and test the algorithm generalization performance on the last fold as detailed below.

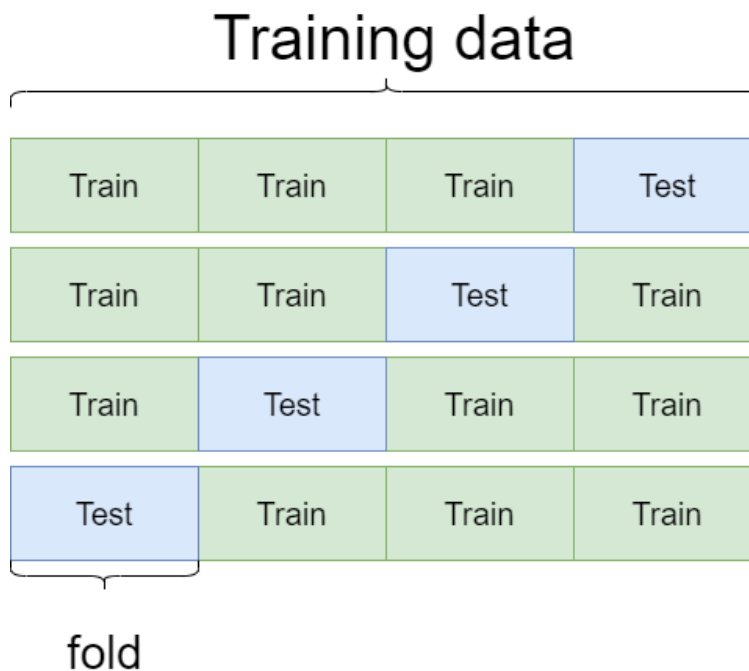


Figure 3.13 example of 4 folds cross validation

Each row returns the accuracy of the model on its testing part. This can be used to check for model robustness as well.

The metric used was the accuracy of the test portion of the cross-validation. The tested models were K-Nearest neighbors, Extra Trees, Random Forest, and Xgboost. They were chosen based on previous literature, [6] [5] and the representation of each available technique (assembling, bagging, and boosting).

3.4.5.1 Deep Learning Training Process

Training is oriented by example and loss through the back propagation as explained in 2.4.2. One of the options to feed the network example is to use a generator which creates the batch for the network to train on. Batches represent stacked inputs. The generator allows for the easier creation of our custom input containing multiple images and clinical data. The optimizer used was Adam [69] starting with $LR = 0.006$, using a step scheduler. Optimizer are algorithms designed to update the weights using the loss value achieved on the previous batch. There are different available optimizers, Adam is a standard optimizer. They also constitute a hyperparameter from the model. The step scheduler is an algorithm modifying the Learning Rate (LR) as the epoch's progress. The steps scheduler allows us to multiply

the LR by a specific coefficient every epoch. The original LR was 0.5 with a minimum of 10^{-6} , so it will never be smaller than this minimum. The model was stopped through early stopping when the validation loss did not decrease for 15 epochs or more. The loss used was categorical cross-entropy, which is calculated as follows.

$$H(p, q) = - \sum_{x \in \mathcal{X}} p(x) \log q(x)$$

The metric used was categorical accuracy in defining improvement of the model.

3.4.6 Optimization of the Problem

Models are designed to generalize from examples, but a problem is not often single-sided. Therefore it is also possible to change the point of view on the issue to try and see if accuracy improves.

3.4.6.1 Regression and Classification

As MJOa represents a continuous target, changing the value to two categorical options might create a loss in information. For example, the difference between a patient getting a 5-point improvement in MJOA after surgery and patient getting only a 2-point improvement seems to be relevant. However, through categorization, we put all these patients on the same baseline. The question is then more complex: is it possible to predict the post-surgical MJOA improvement of patients? This might allow for other features to be learned. The network difference was the final activation layer, as well as the loss used in training. The final activation layer was changed from Softmax, which is created to divide into multiple categories, to linear, which is created for linear regression with a deep learning model. As a loss, we used Mean Squared Error, which computes the square difference between the real mJOA differential and the predicted ones. To measure the accuracy of the network, we converted our predicted MJOA to a binary target using the same discrimination methods that we have used previously ($\Delta M_{joa} \geq 2$). This was done to compare the results with the previous method.

3.4.6.2 Ratio and Minimum Clinical Difference

The recovery ratio as described in [48] will give different answers than our target criteria for severe cases. The $\Delta M_{joa} \geq 2$ as described in 3.1.1 offers a more consistent target. However it could have an impact on the learning process and the final results. This does not change the approach of the problem drastically but it modifies the balance of the dataset which goes from 0.63 to 0.56. The recovery ratio is separated in two categories as well (ratio ≥ 0.5).

3.4.7 Optimization of Models

Models possess multiple optimizable 'hyper parameters.' These parameters are defined at the model creation and aren't modified during the automatic training process. This mostly entails dropout probability, number of layers, bias initialization, optimizer, and learning rate in deep learning. This can be optimized through the various processes. In a more classical machine learning approach, this consists of more algorithm-specific parameters such as tree depth size of the subsample given to each tree, etc.

We mainly focused on two options: Bayesian optimization for our ML model and grid search for our DL models.

3.4.7.1 Grid Search

Grid search is a common method in hyperparameter optimization. The user defines a list of possible values that can be taken by a specific parameter. Once this is done for multiple parameters, the grid search solution is used to test out every possible combination of these parameters. This requires time as each combination needs to be tested.

3.4.7.2 Bayesian Optimization

Bayesian Optimization (BO) is a more advanced solution for the optimization of the hyper parameter. It differs from the grid search process because the process keeps track of its progress and its performance. Grid search just exhausts every possibility asked by the user. The BO process uses its past performance to create a probabilistic model. This probabilistic model is used to determine the optimal hyperparameter for the ML algorithm. The approach is different because the optimization algorithm is given a range of continuous possibilities, which should allow for more configurations to be tested. Moreover, this is a joint optimization of all parameters at the same time [70].

3.5 Evaluation and Metrics

3.5.1 Metrics

For this study the metrics used were accuracy and categorical accuracy as well as Area Under the Curve (AUC). Its baseline is defined by the most represented class in the dataset. Categorical accuracy is calculated, the same way the difference resides in the representation of the output. Categorical outputs, used in deep learning methods, represents a probabilistic point of view. The predicted class will be the most probable. This, however, gives insight

into the results. A good accuracy could come from a lucky, yet indecisive model (e.g., all prediction will be similar to $[0.56, 0.54]$) which would show the inability of the model to learn.

The AUC is a common method of evaluation in a study such as this one. It is computed based on the Receiver Operating Characteristic (ROC). The ROC curves that represent the True Positive Rate (TPR, also known as recall) in regards to the rate of false positive. The rate of true positive is defined as follows

$$TPR = \frac{TP}{TP+FP} \text{ with TP the number of true positive and FP the number of false positive.}$$

3.5.2 Evaluation

For the machine learning model, accuracy was computed based on the mean score obtained on 50 different data split between training and testing (80% and 20% respectively). The model was trained and tested each time. This mitigates the possibility of getting an exceptionally high or low score due to data separation as the dataset size is limited.

For the deep learning model, the accuracy was first measured on the validation set and then on the testing set if the model converged toward what appears a good solution (no overfitting or indecisiveness).

CHAPTER 4 RESULTS

4.0.1 Model Selection

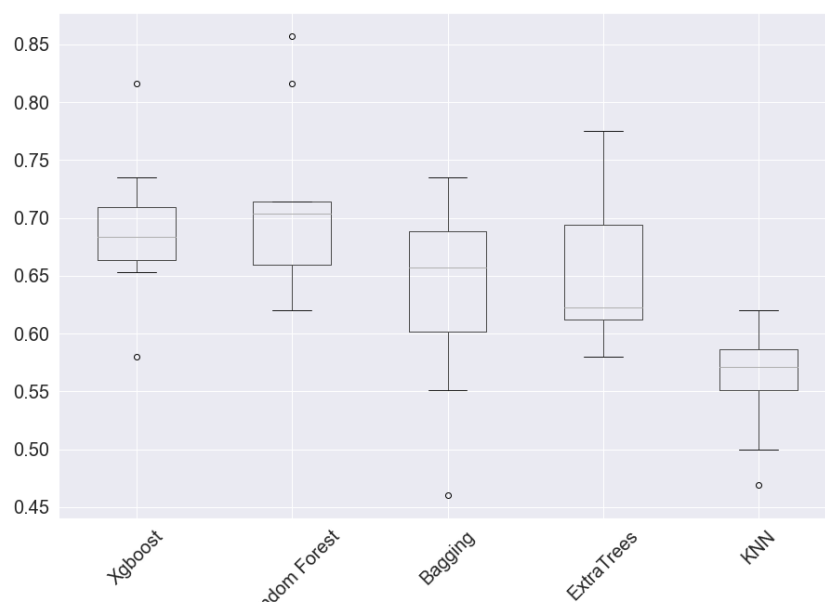


Figure 4.1 *boxplot showing cross-validation scores without operation data*

This was obtained with a 10-fold cross-validation performed by the different model. The best models appear to be Random Forest and Xgboost. This is on par with the previous studies. The boosting technique appears to be less accurate than Random Forest yet, it seems relevant to include it as it looks more consistent, which can be seen by a smaller boxplot. Based on this result, the choice was made to explore these two models.

The models don't seem to provide a steady outcome, as it can be seen through the wide range of accuracy. This is to be expected with a small dataset such as ours. The random split between training and testing fold could influence the outcome.

4.0.2 Surgery Details

This section presents the AUC and accuracy obtained with the chosen models on the data with operation length and operated and without this piece of information.

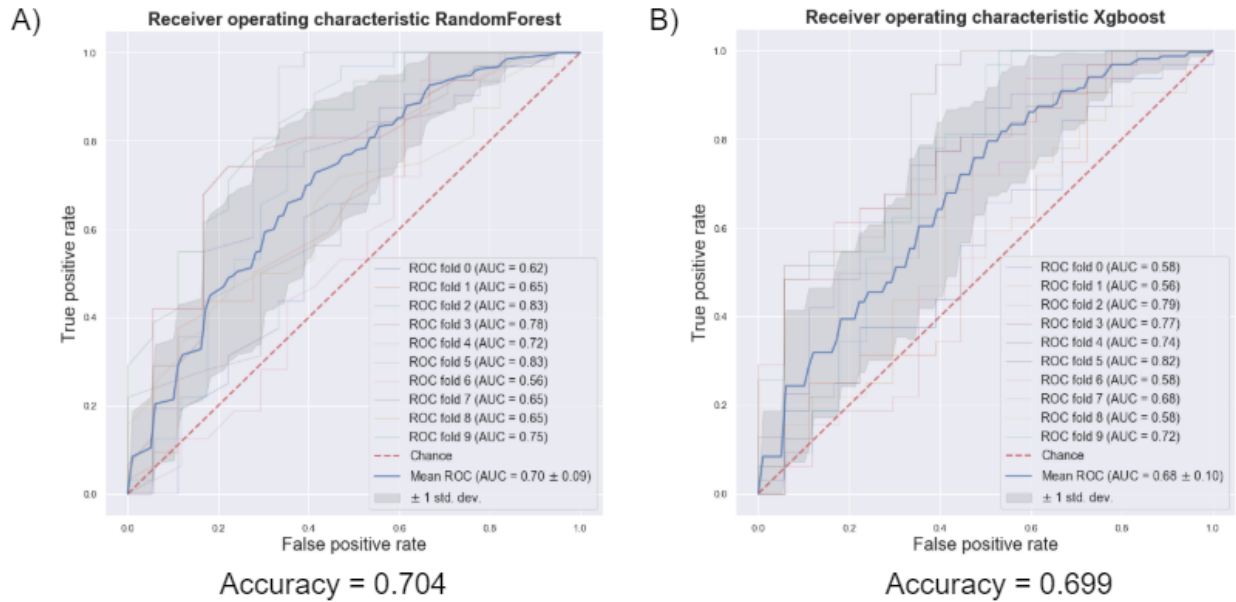


Figure 4.2 *Ten-folds cross-validation AUC with operation information*

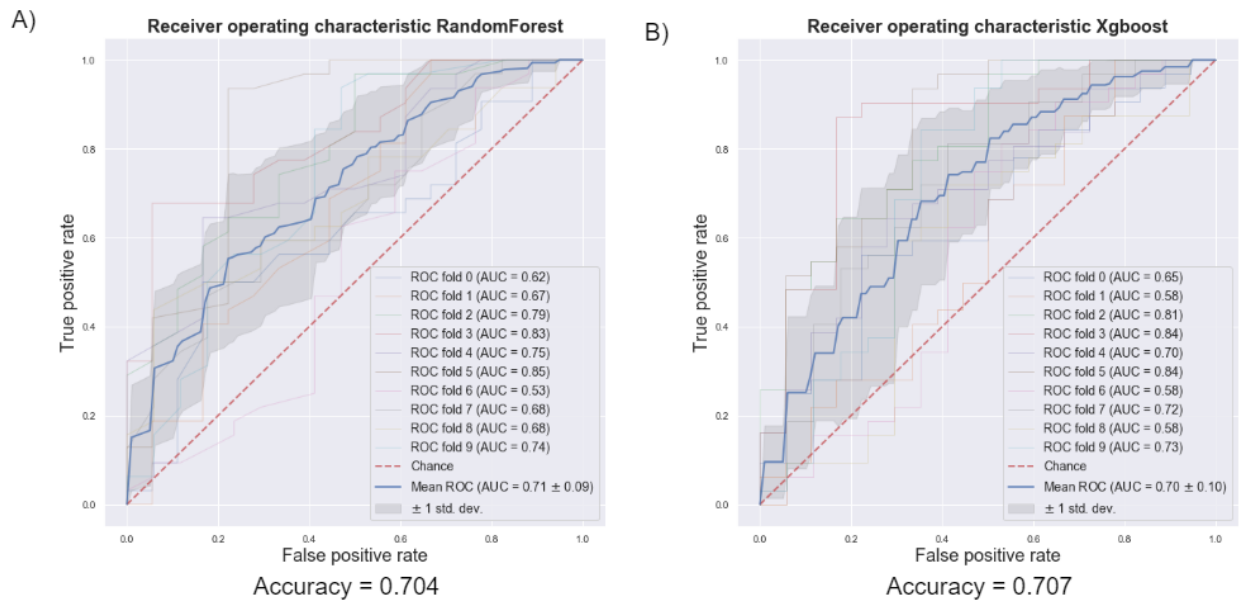


Figure 4.3 *Ten-folds cross-validation AUC without operation information*

There is not much change in the performance of the model considering the availability of this information or not. This data had yet a certain importance according to the model first estimation that can be seen here.

The difference in accuracy is not significant between the two datasets for each model

(p -value >0.05). The slight AUC variation is probably linked to the change in accuracy. As these were included in previous study, it seemed relevant to mention their influences on the results. This is important as this helps to move toward a real pre-surgery prognosis. However, information about the approach and types of the operation remain. It could be interesting to focus on a single surgeon and surgery type to see if it could increase the model performance by removing some variables.

4.0.3 Bayesian Optimization

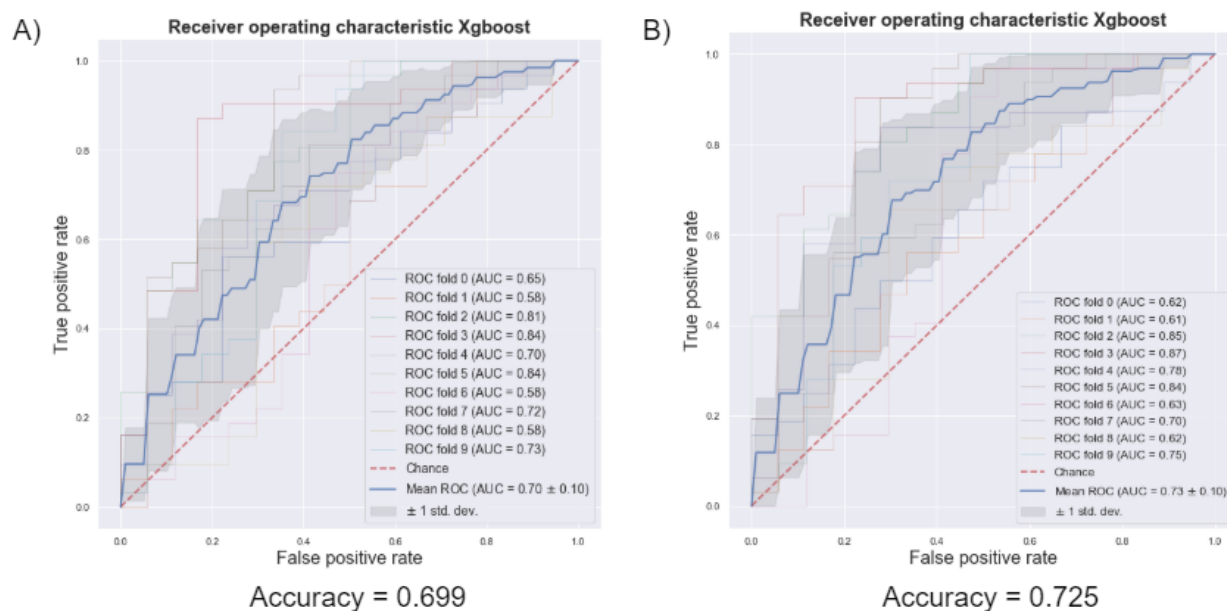


Figure 4.4 A) Ten-folds cross-validation AUC without Bayesian optimization using Xgboost. B) Ten-folds cross-validation AUC with Bayesian optimization using the Xgboost model.

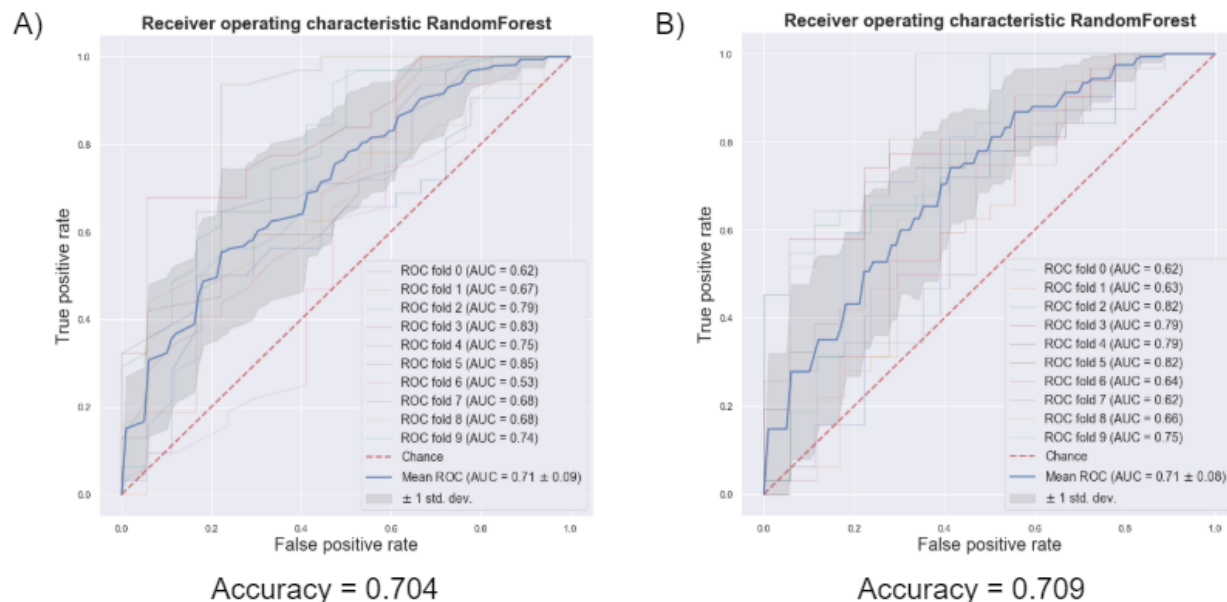


Figure 4.5 A) *Ten-fold cross-validation AUC without Bayesian optimization using the Random Forest model.* B) *Ten-folds cross-validation AUC with Bayesian optimization using the Random Forest model*

The optimization provides an improvement in accuracy and AUC, which is expected because the base parameter had a rare chance of being adapted to the data. This improvement was reflected in the accuracy by an improvement from 69.9% to 72.5% for Xgboost, and from 70.4% to 70.9% for Random Forest.

The difference between the optimized model and the default one was significant for Xgboost (p -value = 0.012). This was not the case for the random forest model. This is interesting as not many studies use BO, though it is more time efficient than grid search. Moreover, this shows that the model will be sensitive to the variation of hyperparameters.

4.0.4 Ratio and MCID

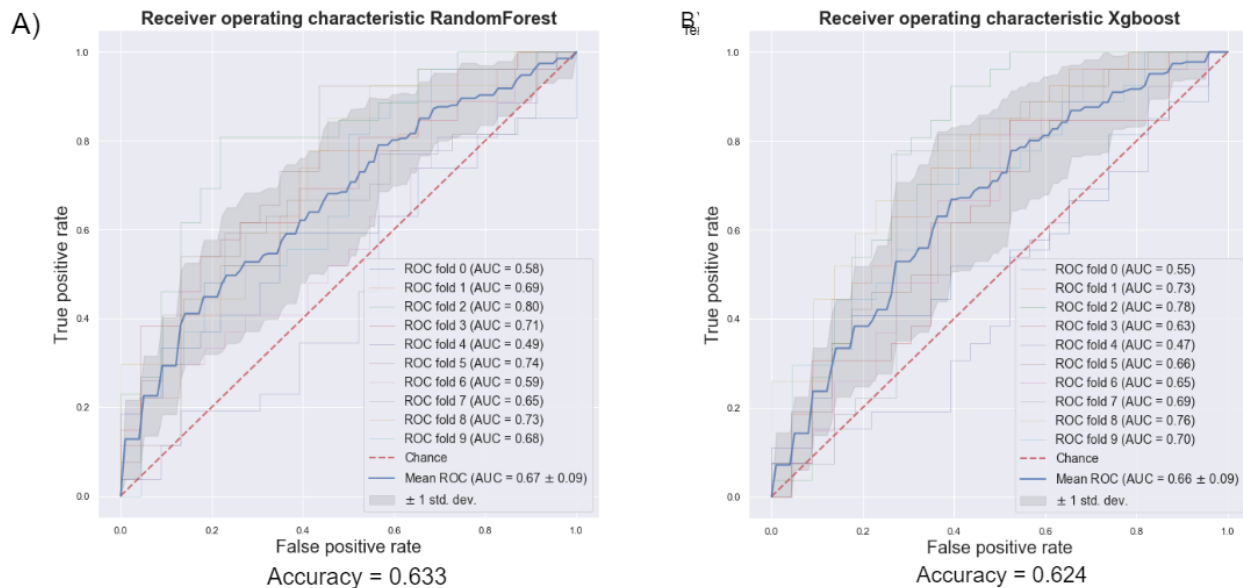


Figure 4.6 Ten-folds cross-validation on the dataset with recovery ratio as ground truth.

Accuracy must be interpreted in regard to the baseline. The baseline for the figures 4.2 was 0.640, while for the results presented above it is only 0.535. Both results show a significant improvement compared to the baseline situation. The improvement obtained from this ratio is, however, slightly greater.

We still remain close to a 9% improvement in accuracy but the AUC appears to be slightly better than in our previous results (in comparison with the baseline). This could be due to the fact that the minimum clinically significant difference commonly used is 2, but it should actually be slightly higher for more severe cases and slightly lower for more mild cases [8]. This could be interpreted as the fact that the ratio provides a more coherent target, but it still relies on the physical examination and therefore could be biased. Moreover, the ratio considers a more "case specific" value, which seems and is more coherent for mild cases (pre-operative Mjoa ≥ 15), because patients can get close to a perfect score after the surgery and still get a $\Delta Mjoa < 2$ due to their high pre-surgery score. However, this ratio seems less relevant on the more severe cases since the patient needs to retrieve more than 50% of his sensorimotor skills, which is a daunting task. In that case the minimum difference seems to be more coherent to at least acknowledge the fact that the surgery helped him preserve some of his sensorimotor skills.

4.0.5 Deep Learning

The Resnet network was fed T2 images only after resampling, cropping, and straightening to try and normalize the input image. Neural networks are using batches of same-sized images as input, which was helped by using crop methods.

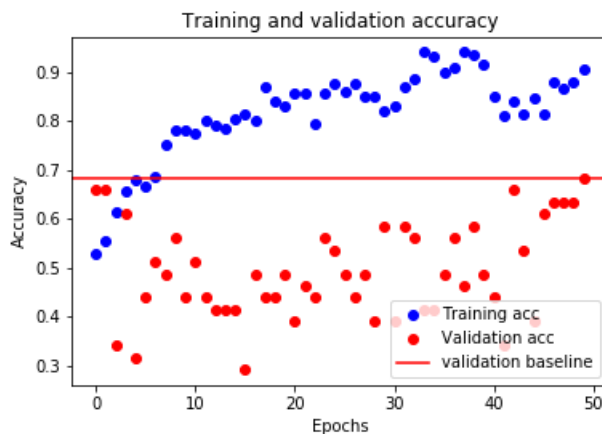


Figure 4.7 ResNet-29 accuracy during training process

There seem to be a clear overfitting process during the training. The validation accuracy remains below the baseline which was 0.6829 after separation in validation and training set. The first Resnet result (figure 4.8) indicates a clear overfitting with 30 layers. This tends to show that the model is too complex for the data that we have: it will be easier for it to memorize every example rather than trying to find a pattern. This was already mitigated through the use of heavy dropout ($p > 0.6$) and Batch Normalization in the classifier step. This has been further mitigated by performing data augmentation. The validation accuracy let transpired that the network is only returning ones or zeros. The loss was decreasing, which is explained by the fact that most predictions were $[0.54, 0.56]$, which represents the indecisiveness of the network.

The Multi-input network was underfitting the data and provided no improvements from the baseline. It was in a similar situation as the Resnet with an important indecisiveness.

These underwhelming performances could be explained by different issues inherent to our problem and our dataset. First, the target might not be fully coherent from patient to patient. Indeed, as explained in 3.1.4, the score can depend on the state of the patient on the day he was evaluated and the bias of the investigator. As the investigator may vary between patient, this could partially explain the underperformance of this approach. Moreover, the

data may differ from patient to patient. As mentioned in 3.1.1, the MR acquisition protocol was not defined which explains a variation in image contrasts and quality. Though it allows to create a credible real-life scenario, it could also hinder the model performance, considering the low number of available images to begin with.

The regularization methods described in 3.4.2 did not have a huge impact on the model performance. Data augmentation did not help with the underfitting and overfitting problem. This might be due to the limited transformation selected. Transformed images might still be too close to the original to be considered "new" by the network. Dropout seems to have an effect but this was too strong at some point because the model went for overfitting to underfitting due to the dropout with a high probability that we try to use. Unfortunately dropout with a weaker probability was not very useful either as the model still overfit our data.

4.0.6 Atlas Analysis

The atlas provides new features as presented in 3.3.5.2, which help to complete the database. However, due to data availability this processing is only performed on 148 subjects. This reduces the baseline to 57.5% of positive cases.

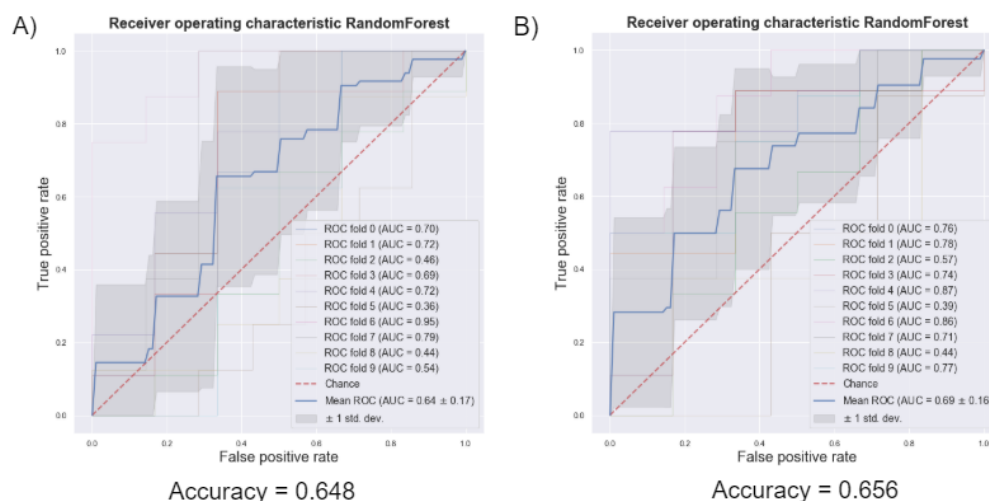


Figure 4.8 AUC obtained on the dataset completed with atlas features, using the minimum clinically significant difference as ground truth for the training. A) Results obtained with Random Forest. B) Results obtained with Xgboost

The improvement from the baseline is significant for both models (p -value > 0.5). It is lower than what can be seen in the other results displayed. This could be influenced by the

patient selection. The number of subjects is fairly low, yet the pipeline is interesting and could help spawn future research in the field. The number of labeled tracts might hinder the model as well, as some information could be less relevant than others. This could be explored in future research with new technologies such as explainable AI, in order to discover which regions in the WM need to be considered in future studies.

CHAPTER 5 CONCLUSION AND RECOMMENDATION

5.1 Summary of Works

This work presents a new approach for the prognosis of degenerative cervical prognosis with automatic machine learning methods. The study evaluates multiple approaches to try and improve the prognosis value on available data from a prospective study: AOspine [55]. Multiple preprocessing was done to improve the chance of getting better results. The goal was to focus on important aspects used now by clinicians and found in the literature.

The study shows that state-of-the-art deep learning models are not performing well on getting information from images and clinical scores on our dataset. Feature extraction techniques and the use of machine learning model appears to have a better chance of providing a solution. The use of MRI paired with the clinical data seems to improve the accuracy by providing new features that can complete the current clinical data obtained with the patient's medical history and the evaluation of the condition.

This could help create a semi-automatic pipeline using the spinal cord toolbox on structural images and clinical information to help decision-making for surgeons in DCM cases.

5.2 Future Research

5.2.1 Surgery Decision

One of the future research could be to focus on the anterior approach for the surgery. The risk of such operation differs from a posterior approach surgery. It entails fewer changes that might affect the patient's mJOA score which could provide a more consistent target. [15] It can be less efficient in some cases. This could help the model by removing some "random" factors in the surgery but will present only cases that require this type of surgery which will create a bias.

5.2.2 DWI

The DWI (or DTI) approach for the clinical evaluation of the patient affected by DCM seems to recently gain interest [71] [72]. This could affect the new study aiming at making post-surgical prognosis and improves on existing ideas as the one presented by Jin et al. [48]. The use of such database paired with an atlas as complete as the one used in this study

might provide interesting results by giving more features to the model. This could give more insight about the effective physiologic state of the patient. This study and others also helped presents different features of importance that would be of use in future study. [5]

REFERENCES

- [1] A. Nouri *et al.*, “Degenerative Cervical Myelopathy: Epidemiology, Genetics, and Pathogenesis,” *Spine*, vol. 40, no. 12, p. E675, Jun. 2015. [Online]. Available: https://journals.lww.com/spinejournal/fulltext/2015/06150/Degenerative_Cervical_Myelopathy___Epidemiology,.8.aspx
- [2] J. H. Badhiwala and J. R. Wilson, “The Natural History of Degenerative Cervical Myelopathy,” *Neurosurgery Clinics*, vol. 29, no. 1, pp. 21–32, Jan. 2018, publisher: Elsevier. [Online]. Available: [https://www.neurosurgery.theclinics.com/article/S1042-3680\(17\)30098-0/abstract](https://www.neurosurgery.theclinics.com/article/S1042-3680(17)30098-0/abstract)
- [3] S. Kato and M. Fehlings, “Degenerative cervical myelopathy,” *Current Reviews in Musculoskeletal Medicine*, vol. 9, no. 3, pp. 263–271, Jun. 2016. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4958380/>
- [4] B. Hilton *et al.*, “Route to diagnosis of degenerative cervical myelopathy in a UK healthcare system: a retrospective cohort study,” *BMJ Open*, vol. 9, no. 5, p. e027000, May 2019. [Online]. Available: <https://bmjopen.bmj.com/content/9/5/e027000>
- [5] Z. G. Merali *et al.*, “Using a machine learning approach to predict outcome after surgery for degenerative cervical myelopathy,” *PLoS ONE*, vol. 14, no. 4, Apr. 2019. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6448910/>
- [6] O. Khan *et al.*, “Machine Learning Algorithms for Prediction of Health-Related Quality-of-Life after Surgery for Mild Degenerative Cervical Myelopathy,” *The Spine Journal*, Feb. 2020. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1529943020300474>
- [7] S. Kato *et al.*, “Comparison of Anterior and Posterior Surgery for Degenerative Cervical Myelopathy: An MRI-Based Propensity-Score-Matched Analysis Using Data from the Prospective Multicenter AOSpine CSM North America and International Studies,” *The Journal of Bone and Joint Surgery. American Volume*, vol. 99, no. 12, pp. 1013–1021, Jun. 2017.
- [8] L. Tetreault *et al.*, “The Minimum Clinically Important Difference of the Modified Japanese Orthopaedic Association Scale in Patients with Degenerative Cervical Myelopathy,” *Spine*, vol. 40, no. 21, pp. 1653–1659, Nov. 2015.

- [9] B. De Leener *et al.*, “SCT: Spinal Cord Toolbox, an open-source software for processing spinal cord MRI data,” *NeuroImage*, vol. 145, no. Pt A, pp. 24–43, 2017.
- [10] A. L. Asher *et al.*, “Comparison of Outcomes Following Anterior vs Posterior Fusion Surgery for Patients With Degenerative Cervical Myelopathy: An Analysis From Quality Outcomes Database,” *Neurosurgery*, vol. 84, no. 4, pp. 919–926, 2019.
- [11] J. Milligan *et al.*, “Degenerative cervical myelopathy,” *Canadian Family Physician*, vol. 65, no. 9, pp. 619–624, Sep. 2019. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6741789/>
- [12] L. Rouhier *et al.*, “Spine intervertebral disc labeling using a fully convolutional redundant counting model,” *arXiv:2003.04387 [cs, eess]*, Mar. 2020, arXiv: 2003.04387. [Online]. Available: <http://arxiv.org/abs/2003.04387>
- [13] A. R. Martin *et al.*, “Monitoring for myelopathic progression with multiparametric quantitative MRI,” *PloS One*, vol. 13, no. 4, p. e0195733, 2018.
- [14] S. Kato, M. Ganau, and M. G. Fehlings, “Surgical decision-making in degenerative cervical myelopathy – Anterior versus posterior approach,” *Journal of Clinical Neuroscience*, vol. 58, pp. 7–12, Dec. 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0967586818307252>
- [15] F. Shou *et al.*, “Prevalence of C5 nerve root palsy after cervical decompressive surgery: a meta-analysis,” *European Spine Journal*, vol. 24, no. 12, pp. 2724–2734, Dec. 2015. [Online]. Available: <https://doi.org/10.1007/s00586-015-4186-5>
- [16] S. Jacobson and E. M. Marcus, *Neuroanatomy for the Neuroscientist*, 2nd ed. Springer US, 2011. [Online]. Available: <https://www.springer.com/gp/book/9781489991348>
- [17] E. R. Kandel, *Principles of neural science*, 5th ed. New York: McGraw-Hill Medical, 2012. [Online]. Available: <http://lib.myilibrary.com?id=396874>
- [18] E. P. W. Dr, H. Raff, and K. T. S. Dr, *Vander’s Human Physiology: The Mechanisms of Body Function*, 13rd ed. New York: McGraw-Hill Education, Mar. 2013.
- [19] S. Lévy *et al.*, “White matter atlas of the human spinal cord with estimation of partial volume effect,” *NeuroImage*, vol. 119, pp. 262–271, Oct. 2015.
- [20] S. Rossignol, “Plasticity of connections underlying locomotor recovery after central and/or peripheral lesions in the adult mammals,” *Philosophical Transactions of the*

Royal Society of London. Series B, Biological Sciences, vol. 361, no. 1473, pp. 1647–1671, Sep. 2006.

- [21] “MRI Basics.” [Online]. Available: <https://casemed.case.edu/clerkships/neurology/Web%20Neurorad/MRI%20Basics.htm>
- [22] C. Gros *et al.*, “Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks,” *NeuroImage*, vol. 184, pp. 901–915, Jan. 2019. [Online]. Available: <https://linkinghub.elsevier.com/retrieve/pii/S1053811918319578>
- [23] D. Eden *et al.*, “Spatial distribution of multiple sclerosis lesions in the cervical spinal cord,” *Brain*, vol. 142, no. 3, pp. 633–646, Mar. 2019, publisher: Oxford Academic. [Online]. Available: <https://academic.oup.com/brain/article/142/3/633/5304670>
- [24] L. J. O’Donnell and C.-F. Westin, “An introduction to diffusion tensor image analysis,” *Neurosurgery clinics of North America*, vol. 22, no. 2, pp. 185–viii, Apr. 2011. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3163395/>
- [25] P. Stroman *et al.*, “The current state-of-the-art of spinal cord imaging: Methods,” *NeuroImage*, vol. 84, pp. 1070–1081, Jan. 2014. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4371133/>
- [26] B. M. Davies *et al.*, “Degenerative cervical myelopathy,” *BMJ*, vol. 360, Feb. 2018, publisher: British Medical Journal Publishing Group Section: Practice. [Online]. Available: <https://www.bmj.com/content/360/bmj.k186>
- [27] S.-A. Park, “Management of elderly patients with spinal disease: Interventional non-surgical treatment,” *Journal of the Korean Orthopaedic Association*, vol. 54, p. 9, 01 2019.
- [28] “Spinal decompression surgery | Cincinnati, OH Mayfield Brain & Spine.” [Online]. Available: <https://mayfieldclinic.com/pe-decompression.htm>
- [29] L. A. Tetreault, A. Karpova, and M. G. Fehlings, “Predictors of outcome in patients with degenerative cervical spondylotic myelopathy undergoing surgical treatment: results of a systematic review,” *European Spine Journal*, vol. 24, no. 2, pp. 236–251, Apr. 2015. [Online]. Available: <https://doi.org/10.1007/s00586-013-2658-z>
- [30] M. G. Fehlings, “Current Knowledge in Degenerative Cervical Myelopathy,” *Neurosurgery Clinics of North America*, vol. 29, no. 1, pp. xiii–xiv, 2018.

- [31] R. Severino, A. Nouri, and E. Tessitore, “Degenerative Cervical Myelopathy: How to Identify the Best Responders to Surgery?” *Journal of Clinical Medicine*, vol. 9, no. 3, Mar. 2020. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7141260/>
- [32] K. Janocha and W. M. Czarnecki, “On Loss Functions for Deep Neural Networks in Classification,” *arXiv:1702.05659 [cs]*, Feb. 2017, arXiv: 1702.05659. [Online]. Available: <http://arxiv.org/abs/1702.05659>
- [33] J. Goschenhofer *et al.*, “Wearable-based Parkinson’s Disease Severity Monitoring using Deep Learning,” *arXiv:1904.10829 [cs, stat]*, Apr. 2019, arXiv: 1904.10829. [Online]. Available: <http://arxiv.org/abs/1904.10829>
- [34] A. S. Lundervold and A. Lundervold, “An overview of deep learning in medical imaging focusing on MRI,” *Zeitschrift für Medizinische Physik*, vol. 29, no. 2, pp. 102–127, May 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0939388918301181>
- [35] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” *arXiv:1505.04597 [cs]*, May 2015, arXiv: 1505.04597. [Online]. Available: <http://arxiv.org/abs/1505.04597>
- [36] M. Havaei *et al.*, “Brain tumor segmentation with Deep Neural Networks,” *Medical Image Analysis*, vol. 35, pp. 18–31, Jan. 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1361841516300330>
- [37] O. Oktay *et al.*, “Attention U-Net: Learning Where to Look for the Pancreas,” *arXiv:1804.03999 [cs]*, May 2018, arXiv: 1804.03999. [Online]. Available: <http://arxiv.org/abs/1804.03999>
- [38] M. Drozdal *et al.*, “The Importance of Skip Connections in Biomedical Image Segmentation,” *arXiv:1608.04117 [cs]*, Sep. 2016, arXiv: 1608.04117. [Online]. Available: <http://arxiv.org/abs/1608.04117>
- [39] S. M. McKinney *et al.*, “International evaluation of an AI system for breast cancer screening,” *Nature*, vol. 577, no. 7788, pp. 89–94, Jan. 2020, number: 7788 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41586-019-1799-6>
- [40] L. Shen *et al.*, “Deep Learning to Improve Breast Cancer Detection on Screening Mammography,” *Scientific Reports*, vol. 9, no. 1, p. 12495, Aug.

- 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41598-019-48995-4>
- [41] Z. Han *et al.*, “Automated Pathogenesis-Based Diagnosis of Lumbar Neural Foraminal Stenosis via Deep Multiscale Multitask Learning,” *Neuroinformatics*, vol. 16, no. 3, pp. 325–337, Oct. 2018. [Online]. Available: <https://doi.org/10.1007/s12021-018-9365-1>
- [42] D. W. Kim *et al.*, “Deep learning-based survival prediction of oral cancer patients,” *Scientific Reports*, vol. 9, no. 1, p. 6994, May 2019, number: 1 Publisher: Nature Publishing Group. [Online]. Available: <https://www.nature.com/articles/s41598-019-43372-7>
- [43] W. Zhu *et al.*, “The Application of Deep Learning in Cancer Prognosis Prediction,” *Cancers*, vol. 12, no. 3, Mar. 2020.
- [44] C.-L. Chi, W. N. Street, and W. H. Wolberg, “Application of Artificial Neural Network-Based Survival Analysis on Two Breast Cancer Datasets,” *AMIA Annual Symposium Proceedings*, vol. 2007, pp. 130–134, 2007. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2813661/>
- [45] P. Mobadersany *et al.*, “Predicting cancer outcomes from histology and genomics using convolutional networks,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 13, pp. E2970–E2979, Mar. 2018, publisher: National Academy of Sciences Section: PNAS Plus. [Online]. Available: <https://www.pnas.org/content/115/13/E2970>
- [46] T. Mizutani *et al.*, “Optimization of treatment strategy by using a machine learning model to predict survival time of patients with malignant glioma after radiotherapy,” *Journal of Radiation Research*, vol. 60, no. 6, pp. 818–824, Nov. 2019.
- [47] B. S. Hopkins *et al.*, “Machine Learning for the Prediction of Cervical Spondylotic Myelopathy: A Post Hoc Pilot Study of 28 Participants,” *World Neurosurgery*, vol. 127, pp. e436–e442, Jul. 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1878875019308459>
- [48] R. Jin *et al.*, “A machine learning based prognostic prediction of cervical myelopathy using diffusion tensor imaging,” in *2016 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications (CIVEMSA)*, Jun. 2016, pp. 1–4, iSSN: 2377-9322.
- [49] B. D. Leener *et al.*, “Topologically preserving straightening of spinal cord MRI,” *Journal of magnetic resonance imaging : JMRI*, 2017.

- [50] “neuropoly/sct_docker,” Feb. 2020, original-date: 2018-02-15T21:11:37Z. [Online]. Available: https://github.com/neuropoly/sct_docker
- [51] E. Ullmann *et al.*, “Automatic Labeling of Vertebral Levels Using a Robust Template-Based Approach,” *International Journal of Biomedical Imaging*, vol. 2014, 2014. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4123554/>
- [52] J. P. Cohen *et al.*, “Count-ception: Counting by Fully Convolutional Redundant Counting,” *arXiv:1703.08710 [cs, stat]*, Jul. 2017, arXiv: 1703.08710. [Online]. Available: <http://arxiv.org/abs/1703.08710>
- [53] J. Cohen-Adad, “Spinal Cord MRI Public Database (Multi-subjects) - Snapshot 1.0.4 - OpenNeuro,” Jul. 2019. [Online]. Available: <https://openneuro.org/datasets/ds001919/versions/1.0.4>
- [54] B. Kopjar *et al.*, “The AOSpine North America Cervical Spondylotic Myelopathy Study: Perioperative Complication Rates Associated with Surgical Treatment Based on a Prospective Multicenter Study of 302 Patients,” *The Spine Journal*, vol. 11, no. 10, p. S85, Oct. 2011, publisher: Elsevier. [Online]. Available: [https://www.thespinejournalonline.com/article/S1529-9430\(11\)00735-2/abstract](https://www.thespinejournalonline.com/article/S1529-9430(11)00735-2/abstract)
- [55] M. G. Fehlings *et al.*, “A global perspective on the outcomes of surgical decompression in patients with cervical spondylotic myelopathy: results from the prospective multicenter AOSpine international study on 479 patients,” *Spine*, vol. 40, no. 17, pp. 1322–1328, Sep. 2015.
- [56] K. He *et al.*, “Deep Residual Learning for Image Recognition,” *arXiv:1512.03385 [cs]*, Dec. 2015, arXiv: 1512.03385. [Online]. Available: <http://arxiv.org/abs/1512.03385>
- [57] C. Szegedy *et al.*, “Going Deeper with Convolutions,” *arXiv:1409.4842 [cs]*, Sep. 2014, arXiv: 1409.4842. [Online]. Available: <http://arxiv.org/abs/1409.4842>
- [58] B. De Leener *et al.*, “PAM50: Unbiased multimodal template of the brainstem and spinal cord aligned with the ICBM152 space,” *NeuroImage*, vol. 165, pp. 170–179, Jan. 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1053811917308686>
- [59] “Spinal Cord Toolbox / Documentation / registration_tricks.” [Online]. Available: https://sourceforge.net/p/spinalcordtoolbox/wiki/registration_tricks/

- [60] F. Pedregosa *et al.*, “Scikit-learn: Machine Learning in Python,” *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: <http://jmlr.org/papers/v12/pedregosa11a.html>
- [61] L. Breiman, “Random Forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, Oct. 2001. [Online]. Available: <https://doi.org/10.1023/A:1010933404324>
- [62] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, Aug. 2016, arXiv: 1603.02754. [Online]. Available: <http://arxiv.org/abs/1603.02754>
- [63] X. Ji *et al.*, “Five-Feature Model for Developing the Classifier for Synergistic vs. Antagonistic Drug Combinations Built by XGBoost,” *Frontiers in Genetics*, vol. 10, 2019, publisher: Frontiers. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fgene.2019.00600/full>
- [64] N. Srivastava *et al.*, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014. [Online]. Available: <http://jmlr.org/papers/v15/srivastava14a.html>
- [65] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” *arXiv:1502.03167 [cs]*, Mar. 2015, arXiv: 1502.03167. [Online]. Available: <http://arxiv.org/abs/1502.03167>
- [66] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016, <http://www.deeplearningbook.org>.
- [67] S. Guo and Z. Yang, “Multi-Channel-ResNet: An integration framework towards skin lesion analysis,” *Informatcs in Medicine Unlocked*, vol. 12, pp. 67–74, Jan. 2018. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S2352914818300868>
- [68] A. Nouri *et al.*, “Magnetic resonance imaging assessment of degenerative cervical myelopathy: a review of structural changes and measurement techniques,” *Neurosurgical Focus*, vol. 40, no. 6, p. E5, Jun. 2016.
- [69] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Jan. 2017, arXiv: 1412.6980. [Online]. Available: <http://arxiv.org/abs/1412.6980>

- [70] J. Snoek, H. Larochelle, and R. P. Adams, “Practical Bayesian Optimization of Machine Learning Algorithms,” in *Advances in Neural Information Processing Systems 25*, F. Pereira *et al.*, Eds. Curran Associates, Inc., 2012, pp. 2951–2959. [Online]. Available: <http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf>
- [71] A. Nouri *et al.*, “Degenerative Cervical Myelopathy: A Brief Review of Past Perspectives, Present Developments, and Future Directions,” *Journal of Clinical Medicine*, vol. 9, no. 2, Feb. 2020. [Online]. Available: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7073521/>
- [72] J. H. Badhiwala *et al.*, “Degenerative cervical myelopathy - update and future directions,” *Nature Reviews. Neurology*, vol. 16, no. 2, pp. 108–124, 2020.