College of Sciences and Technologies in Engineering – FCTE/UnB
Biomedical Engineering Graduate Program

Proposal and Evaluation of Methods for the Segmentation of the Brainstem and of Regions Related to Parkinson Disease's Staging, based on U-Nets applied to Magnetic Resonance Images

Gabriela Kaori Diógenes

Advisor: Cristiano Jacques Miosso Rodrigues Mendes

University of Brasilia

College of Sciences and Technologies in Engineering

Proposal and Evaluation of Methods for the Segmentation of the Brainstem and of Regions Related to Parkinson Disease's Staging, based on U-Nets applied to Magnetic Resonance Images

Gabriela Kaori Diógenes

Advisor: Cristiano Jacques Miosso Rodrigues Mendes

Master Degree Thesis on
Biomedical Engineering

Publication: 209A/2026

Brasilia/DF, janeiro de 2026

University of Brasilia

College of Sciences and Technologies in Engineering

Graduate Program

Proposal and Evaluation of Methods for the Segmentation of the Brainstem and of Regions Related to Parkinson Disease's Staging, based on U-Nets applied to Magnetic Resonance Images

Gabriela Kaori Diógenes

Master Thesis submitted to the Biomedical Engineering Graduate Program, as a partial fulfillment of the requirements for the degree of Master in Biomedical Engineering

Approved by:

Cristiano Jacques Miosso Rodrigues Mendes

(Advisor)

Prof. Dr. Fabiano Araújo Soares

(Internal examiner)

Prof. Dr. Alexsandro Euripedes Ferreira

(External examiner)

Brasilia/DF, janeiro de 2026

Catalog Card

Diógenes, Gabriela Kaori
Proposal and Evaluation of Methods for the Segmentation of the Brainstem and of Regions Related to Parkinson Disease's Staging, based on U-Nets applied to Magnetic Resonance Images.
[Distrito Federal], 2026.
53p., 210 × 297 mm (FCTE/UnB, Mestrado em Engenharia Biomédica, 2026).
Dissertação de Mestrado em Engenharia Biomédica, Faculdade de Ciências e Tecnologias em Engenharia (FCTE), Programa de Pós-Graduação em Engenharia Biomédica.

1. Neuroimaging
2. Computer Vision
3. Parkison's Disease
4. Artificial Intelligence

I. FCTE UnB/UnB.
II. Título (série)

Reference

Diógenes, Gabriela Kaori (2026). Proposal and Evaluation of Methods for the Segmentation of the Brainstem and of Regions Related to Parkinson Disease's Staging, based on U-Nets applied to Magnetic Resonance Images. Master thesis in Biomedical Engineering, Publication 209A/2026, Biomedical Engineering Graduate Program, University of Brasilia at Gama, Brasilia, DF, 53p.

Copyright

---

gkaori@ieee.org

Brasília, DF – Brasil

Proposta e Avaliação de Métodos para a Segmentação do Tronco Encefálico e de Regiões Relacionadas ao Estadiamento da Doença de Parkinson, com base em U-Nets aplicadas a Imagens de Ressonância Magnética

Resumo

Estudos recentes sobre neuroimagem identificaram biomarcadores em imagens de Ressonância Magnética (RM) que podem servir como critérios de suporte para o diagnóstico e estadiamento da Doença de Parkinson (DP). Entre eles, o sinal de Neuromelanina (NM) na região de Substância Negra (SN) e do Locus Coeruleus (LC) mostrou-se particularmente relevante para este propósito. No entanto, análises modernas dessas regiões do tronco encefálico tipicamente baseiam-se em segmentações manuais e avaliações criteriosas feitas por neurologistas e neurorradiologistas, processo complexo e demorado. Paralelamente, a análise computacional de imagens e Inteligência Artificial (AI) também têm tido consideráveis avanços, especialmente no ramo de imageamento médico, onde modelos de U-Net têm alcançado desempenho estado da arte em segmentação de imagens. Ainda assim, poucos estudos têm aplicado tais modelos para segmentação de estruturas do tronco encefálico considerando diferentes planos anatômicos, possivelmente devido à escassez de bases de dados extensas anotadas por especialistas.

Neste trabalho, utilizamos uma base de dados de imagens de RM ponderadas em T1, criada pelo nosso grupo de pesquisa, para treinar modelos U-Net para segmentação automática de tronco encefálico. Essa tarefa é uma etapa crucial para a detecção automática do sinal de NM na SN e no LC. Nosso método proposto parte da geração de máscaras de referência utilizando o programa Freesurfer seguido de curadoria feita por neurologistas e neurorradiologistas. Em seguida, treinamos 4 diferentes modelos e avaliamos os seus desempenhos com base no Coeficiente de Similaridade de Dice (DSC) e na Intersecção sobre União (IoU). 3 modelos foram treinados unicamente com um tipo de corte anatômico, enquanto o outro com todos os 3 tipos de cortes, o axial, o coronal e o sagital. Para seleção de hiperparâmetros, usamos uma busca por grade para escolher otimizadores, taxas de aprendizado e número de filtros na primeira camada do modelo. Além disso, fizemos uma análise preliminar em um modelo cujo treinamento foi feito com cortes axiais e ajuste fino com imagens de RM sensíveis a NM, que apresentaram resultados promissores na tarefa de segmentar o tronco encefálico em exames feitos no espaço de NM.

O modelo com melhor desempenho, treinado com a base de dados maior de cortes coronais, alcançou DSC de 95,34% e um IoU de 93,03%. Em seguida, veio o modelo treinado com cortes axiais (DSC: 93,88%, IoU: 89,17%) seguido do genérico (DSC: 92,42%, IoU: 87,49%). Em contrapartida, o modelo treinado em cortes sagitais obteve um desempenho inferior quando comparado aos demais (DSC: 87,73%, IoU: 81,87%). Esses resultados sugerem que a seleção de cortes não afetou de maneira crítica o desempenho dos modelos. Ainda assim, os modelos baseados em cortes coronais forneceram as segmentações mais confiáveis, nas

condições testadas. Considerando o estudo de caso do modelo baseado em cortes axiais com ajuste fino de imagens em espaço de NM, ele atingiu 84,75% de DSC e 76,26% de IoU.

Como trabalho futuro, almejamos estudar mais profundamente o processo de ajuste fino realizado no modelo aplicado a imagens de NM para que este faça a segmentação da região do mesencéfalo e, posteriormente, a quantificação do sinal da área hipertensa de NM da SN. Com isso, será possível aplicar um modelo classificador, inicialmente binário, que fará a predição entre DP ou grupo controle. Em seguida, aspiramos escalar esse modelo para um modelo de classificação multiclasses considerando os estágios de DP conforme escala. Por fim, planejamos avaliar diferentes estratégias de limiarização e analisar separadamente os 2 hemisférios do cérebro, com o objetivo final de viabilizar ferramentas de detecção automática de biomarcador relacionado à DP.

Palavras-chave: Doença de Parkinson, segmentação do tronco encefálico, substância negra, neuromelanina, imagens de ressonância magnética, diagnóstico assistido por computador.

Abstract

Recent progress in neuroimaging has identified Magnetic Resonance Imaging (MRI) biomarkers that may support the diagnosis and staging of Parkinson's Disease (PD). Among these, the Neuromelanin (NM) in the Substantia Nigra (SN) and Locus Coeruleus (LC) show particular interest. However, current analyses of these brainstem regions typically rely on manual segmentation and evaluation by neurologists and radiologists, a process that is costly and time consuming. In contrast, computational image analysis and Artificial Intelligence (AI) have recently advanced, especially in medical imaging, where U-Nets achieve state-of-the-art segmentation performance. Yet, few studies have applied such models to brainstem structures, taking into consideration the anatomical planes, likely due to limited availability of large, expertly segmented datasets.

In this work, we use a dataset of T1-Weighted (T1W) MRI images developed by our group to train U-Net models for automatic brainstem segmentation. This task represents a crucial first step toward detecting the NM signal of SN and LC. Our proposed method began with scripted segmentation of the brainstem using the Freesurfer software package, followed by expert validation from neurologists and radiologists. We then trained four models and evaluated their performance using the Dice Similarity Coefficient (DSC) and the Intersection over Union (IoU). Three models were trained using slices from single anatomical planes, so, axial, coronal or sagittal, while one used multi-plane slices. We also used grid search across optimizers, numbers of filters in the first layer, and learning rates. Additionally, we made a preliminary analysis of an axial-based model trained with T1W images and fine-tuned with NM scans that provided promising results in the task of segmenting the brainstem in the NM space.

The best-performing model, trained on coronal slices of the larger subset, achieved DSC of 95.34% and IoU of 93.03%. The axial-based followed (DSC: 93.88%, IoU: 89.17%), with the generic next (DSC: 92.42%, IoU: 87.49%). In contrast, the sagittal model underperformed, compared to the other (DSC: 87.73%, IoU: 81.87%). These results suggest that slice selection does not critically affect performance. Still, coronal-based provided the most reliable results for brainstem segmentation, under the tested conditions. Regarding the case study of the axial-based model fine-tuned with NM images, it achieved an average DSC of 84.75% and IoU of 76.26%.

Our next research steps are to study more profoundly the fine-tuned axial-based model applied to the NM acquisitions so that we can segment only the midbrain and, finally, quantify the hyperintense area of nigral NM. With that, we aim to use a classifier that will, at first, produce binary outputs, between PD and Control Group (CG). Later, it will classify among all stages of PD's staging scale. We also plan to evaluate differ-

ent thresholding strategies and separately analyze the two brain hemispheres, aiming to advance automatic tools for PD-related biomarker detection.

Keywords: Parkinson's Disease, Brainstem Segmentation, Substantia Nigra, Neuromelanin, U-Net, Magnetic Resonance Imaging, Computer-Aided Diagnosis.

Contents

List of Tables

List of Frames

List of Figures

List of Nomenclatures and Abbreviations

AD Alzheimer's Disease

ADNI-3 Alzheimer's Disease Neuroimaging Initiative-3

AG-SE-ResNeXt50 Attention Gated Squeeze-and-Excitation Residual Networks (suggesting next dimension) with 50 layers

AI Artificial Intelligence

ALS Amyotrofic Lateral Sclerosis

Adam Adaptative Moment Estimation

AdamW Adaptative Moment Estimation with Decoupled Weight Decay

ANN Artificial Neural Network

ASSD Average Symmetric Surface Distance

AUC Area Under the Receiver Operating Characteristic Curve

CA-Net Comprehensive Attention-based Convolutional Neural Network (CNN)

CAD Computer-Aided Diagnosis

CG Control Group

CIS Clinically Isolated Syndrome

CNS Central Nervous System

CNN Convolutional Neural Network

CT Computerized Tomography

CTE Chronic Traumatic Encephalopathy

DALY Disability-Adjusted Life Years

DC Diencephalon

DL Deep Learning

DSC Dice Similarity Coefficient

EEG Electroencephalography

fMRI Functional MRI

FU-Net Feedback Weighted U-Net

GBD Global Disease Burden

GPU Graphic Processing Unit

HD Hausdorff Distance

HD95 95th Percentile HD

HIC High-Income Countries

IoU Intersection over Union

LB Lewy Bodies

LC Locus Coeruleus

LMIC Low and Middle-Income Countries

MAE Mean Absolute Error

MAPE Mean Absolute Percentage Error

MCI Mild Cognitive Impairment

ML Machine Learning

MLP Multilayer Perceptron

MRI Magnetic Resonance Imaging

MS Multiple Sclerosis

MS-CoRe-U-Net Multi-Scale CoRe-U-Net

MSE Mean Squared Error

NM Neuromelanin

PD Parkinson's Disease

PET Positron Emission Tomography

PSP Progressive Supranuclear Palsy

PTSD Post-Traumatic Stress Disorder

QSM Quantitative Susceptibility Mapping

RBD Rapid Eye Movement (REM) sleep Behavior Disorder

ReLU Rectified Linear Unit

REM Rapid Eye Movement

ResNet-50 Residual Network with 50 layers

ROI Region of Interest

SDI Socio-Demographic Index

SMHD Symmetrised Modified HD

SN Substantia Nigra

T1W T1-Weighted

T2W T2-Weighted

T2* T2-Star-Weighted

1 Introduction

1.1 Parkinson's Disease

Parkinson's Disease (PD) is a progressive movement disorder and the second most common neurodegenerative disease worldwide [51]. It affects not only people's motor skills, but also cognitive and sensory skills as well. It's prevalence is strongly related to the factor of age and, due to the world's life span increase, its incidence tends to rise even more [51]. Death and disability caused by PD are increasing faster than any other neurological disorder worldwide [61]. In fact, according to [22], PD has remained the second leading cause of death among neurological disorders from 1990 to 2021 with an upward trend projected for the next years, as presented in Graph 1.1.

In the case of Brazil's context regarding PD, in 2021, 31.8 million people were 60 years old or older, representing 14.53% of the country's population [22]. The prevalence of PD among this age range in the same year was of 174.4 thousand people, which means that 5 in 1,000 people over 60 years old had PD. Also in 2021, 9.6 thousand people in this age range died due to PD in Brazilian territory. Some factors that contributed to the increasing prevalence are: an aging population, a longer duration of the disorder, and enhanced diagnostic capabilities [11, 51].

It must be taken into account that prevalence and incidence data on PD are inconsistent, specially for Low and Middle-Income Countries (LMIC) an ethnic minorities in High-Income Countries (HIC) due to misdiagnosis, inefficient surveillance systems, lack of awareness and knowledge of the disease, which leads to erroneous perceptions associating PD's symptoms to the process of aging [42, 61]. Still, Peng et al. [42] presented that not only is there an association between countries' Socio-Demographic Index (SDI) and PD, but also a significant positive correlation between them, regarding Global Disease Burden (GBD) data from 2021. Rates of age-standardized PD prevalence, incidence, Disability-Adjusted Life Years (DALY), and mortality all presented a p-value minor than 0.001 when correlated to the SDI of 21 GBD regions for people over 55 years of age [42]. This analysis indicates a significant impact of demographic aging and healthcare accessibility to PD burden [42]. Additionally, another relation evidenced in Peng et al. [42] is the significantly higher rates of mortality, prevalence and incidence of PD in men than

in women, who have 33% lower risk of developing PD.

Despite the motor symptoms, PD may affect a person's gastrointestinal and urinary tracts, mental health, sleeping patterns, olfactory system, and cognition [6, 8]. In fact, if there were no cognitive impairment by the time of the diagnosis, after 1 year, the cumulative incidence of PD-Mild Cognitive Impairment (MCI) is 9.9%, 23% after 3 years, 29% after 5 years and 39% of these cases are converted to dementia in half a decade [8]. Regarding data from GBD of 2021, PD was the 11th disease with higher age-standardized DALY worldwide, the 5th with higher DALY per 100,000 people aged 60 to 79 years, and the 3rd for people over 79 years, among all conditions with neurological health loss [53].

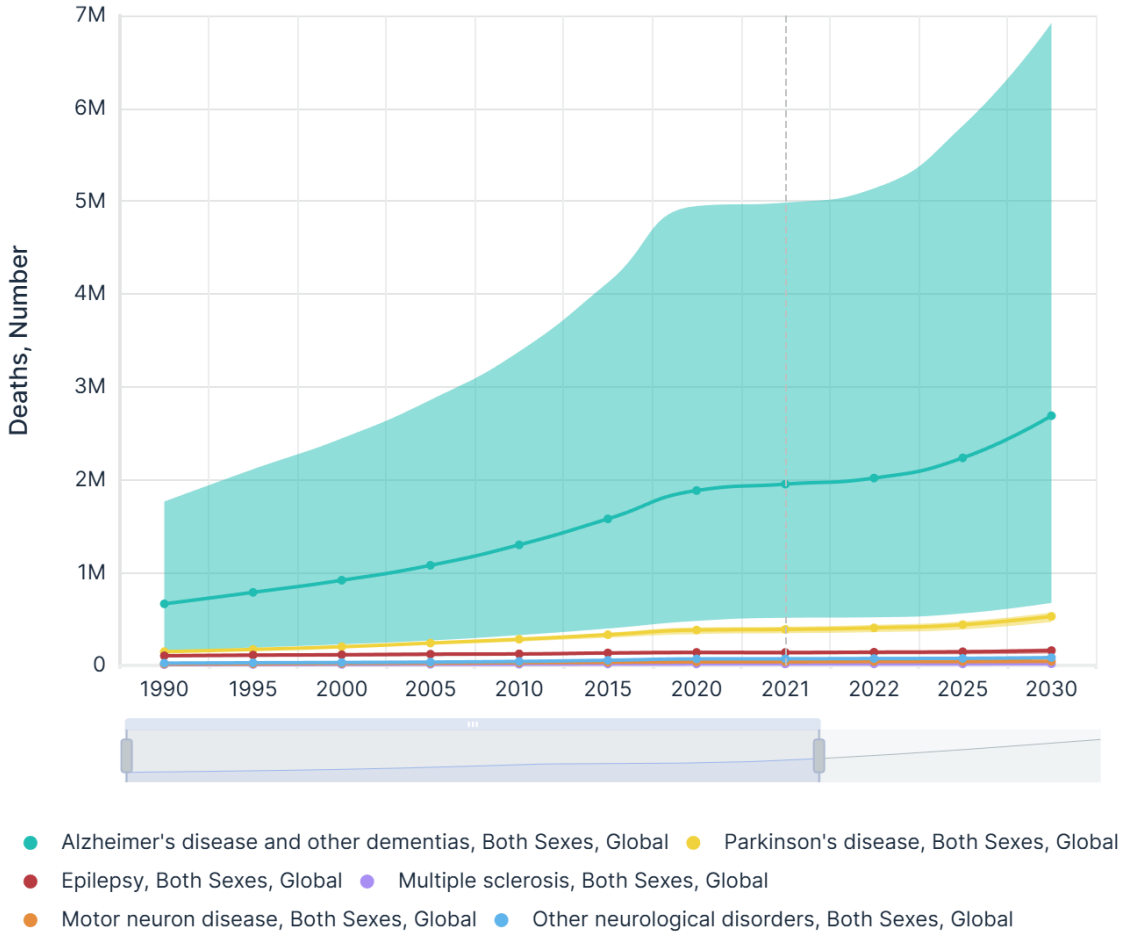


Figure 1.1. Number of deaths caused by neurological disorders between 1990 to 2021 and its projected trend until 2030, considering both sexes and all ages in a global scale. Source: [22].

Considering that there's no treatment that halts the progression of the disease, the earlier the diagnosis is given, the better the prognosis to retard its advancement, because as PD advances, patients tend to become more dependent on care that is mostly provided by informal caregivers [11]. In light of that, PD's burden usually extends strongly to the caregivers, who may experience anxiety or depression, sleep impairment, financial

pressure, and diminished social and leisure time [11].

Only in 2019, more than 35 billion dollars were spent directly with PD globally, more than double the value spent in 2000 [22]. In Brazil, the economic impact was of 354 million dollars [22].

1.2 Neuroimaging and computer vision

Brain imaging was a turning point in the study of brain's functions and structures [5]. Despite its noninvasive aspect, there's a great range of techniques used for specific purposes, such as Electroencephalography (EEG) for electric activity [5, 64], Functional Magnetic Resonance Imaging (MRI) (fMRI) for neural activity, and Positron Emission Tomography (PET) for metabolic activity. In this context, there are 3 fundamental anatomical planes: axial (horizontal), coronal and sagittal, in which medical imaging scans can be acquired according to the visualization needed [2]. These are illustrated in Figure 1.2.

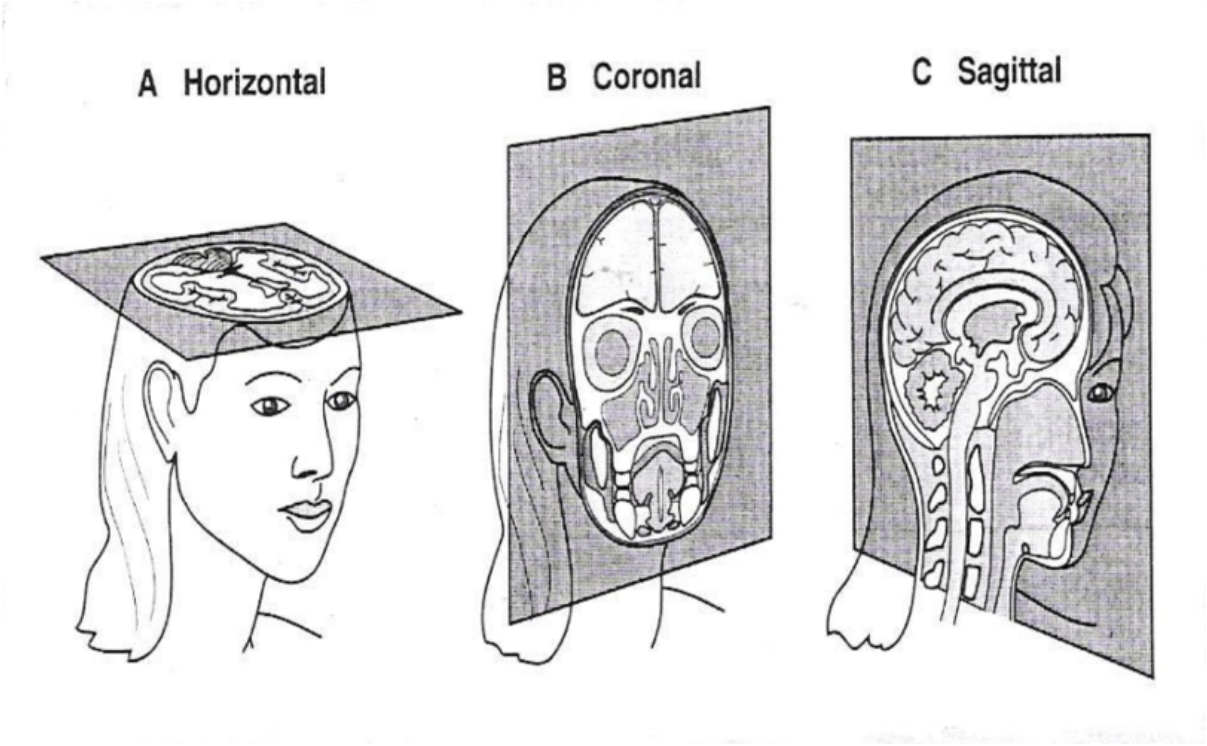


Figure 1.2. Sectional planes used to visualize the human body in medical imaging. To the left, A represents the horizontal, transverse, or axial plane, dividing the body into superior and lower sections. In the center, B illustrates the coronal view, separating the body into anterior and posterior sections. Lastly, to the right, C evidences a sagittal plane, partitioning the body into left and right sections. Source: [44].

MRI is one of the most used imaging techniques due to its characteristic of not using

any type of ionizing radiation, turning it into a much safer option compared to radiological alternatives [5]. Also, the fact that MRI signals come from water molecules within the body, this technique can be applied to a variety of tissues based on their intrinsic contrast parameters, which turns it into a powerful and multifaceted option [5]. When applied to the brain, an example of tissue differentiation can be made between Grey Matter, White Matter and Cerebrospinal Fluid, as shown in Figure 1.3. More specifically, according to [19], T1-Weighted (T1W) scans provide a clearer contrast between the midbrain (part of the brainstem) and the surrounding tissues when compared to T2-Weighted (T2W) and T2-Star-Weighted (T2*), which indicates that it is ideal for defining midbrain contour.

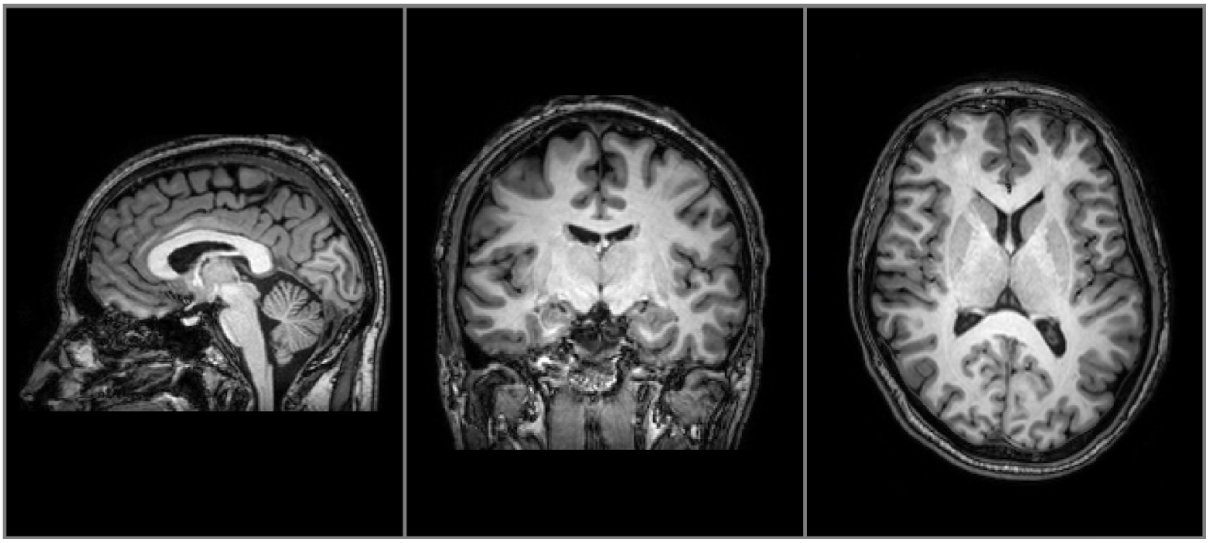


Figure 1.3. T1W image of the brain in sagittal, coronal and axial views, respectively. In this example, cerebrospinal fluid appears dark, gray matter in gray tones and white matter in brighter tones, showing how contrast can differentiate tissues. Source: [5].

Meanwhile, according to [29], computer vision is a subfield of Artificial Intelligence (AI) that tries to imitate the human vision and, for that, both the sensory and cognitive systems are taken into account. The first one relies on capturing images, which nowadays is mainly done by digital cameras [29]. The second concentrates on the extraction of information from the representations given by the first system [29]. Back in 2010, the scenario of computer vision was completely different, once the computer's capability to extract data from images did not involve Machine Learning (ML) methods [29].

One of the main problems addressed with computer vision is segmentation, specifically semantic segmentation, which means classifying each and every pixel of an image [29]. Likewise, neuroimaging segmentation is an important analysis in medical applications once it can identify both normal and abnormal structures in the brain [55]. Therefore, using well-trained ML models in neuroimaging segmentation tasks can be valuable for its ability to detect subtle patterns that may be invisible to the human eye and for its

immunity to human vulnerabilities, such as fatigue [33].

1.3 Scientific Proposal

Considering all the advances done in the last few years related to PD's progression, specially the identification of pre-motor symptoms that may occur decades prior to the appearance of the motor ones, combined with the development of new technologies in diagnostic medicine, the possibilities to discover early stage biomarkers of the disease and even new therapies only raise. So, in the same way that symptoms that were previously disregarded due to the lack of association to PD may now be considered, signals that could not be seen with a reasonable resolution may now be visualized.

At the same time, AI is developing with an extraordinary pace supplied with high processing capacity computers and almost unlimited storage of cloud computing. And one of the AI fields that's being more privileged with these advances is computer vision, exactly because of the high dimensional characteristic of this type of data.

In order to use AI is PD's staging, a crucial step is to be able to segment brain structures related to the disease, specially in images from the Neuromelanin (NM) space. However, scans sensitive to NM usually are less abundant than T1W-MRI, mainly because it focus in a limited field of view and does not cover the whole brain. In our dataset, for example, for each patient, there are 12 slices in NM space and 256 axial slices in T1W. Moreover, the mechanism of signal generation in NM acquisition highlights NM-rich areas, which makes anatomical structures' delineation unreliable. On the other hand, T1W images show clear differentiation between tissues, leading to high anatomical detailing compared not only to NM scans, but also to **T2!** (**T2!**) and T2*, for instance. In Figure 1.4 we evidenced the brainstem's borders in red in both images and the NM signal highlighted is indicated with the yellow arrow.

Thus, it is mostly unfeasible to train a computer vision model only with NM scans due to its insufficiency in terms of quantity and also detailed anatomical differentiation. In this respect, our proposal is to train a model firstly with T1W scans, so it can learn the patterns of the brainstem contour. After that, we aim to make use of transfer learning fundamentals to fine-tune this pre-trained model with NM images.

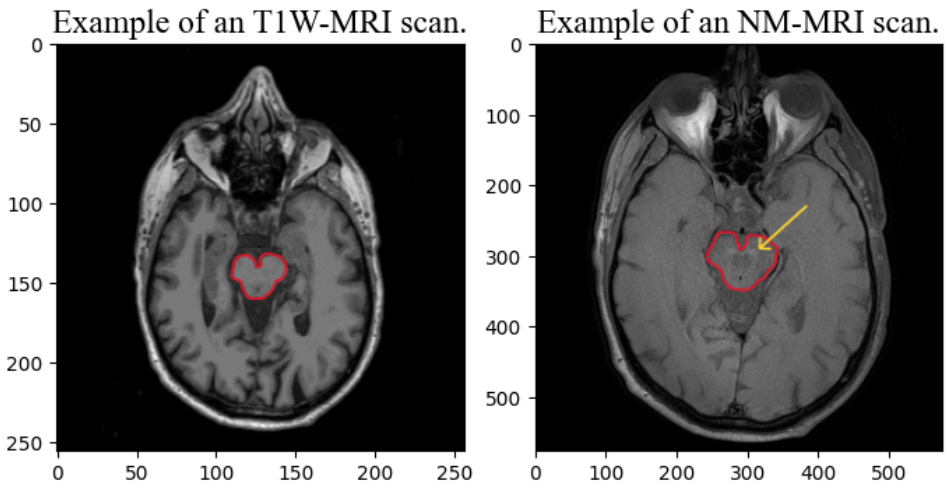
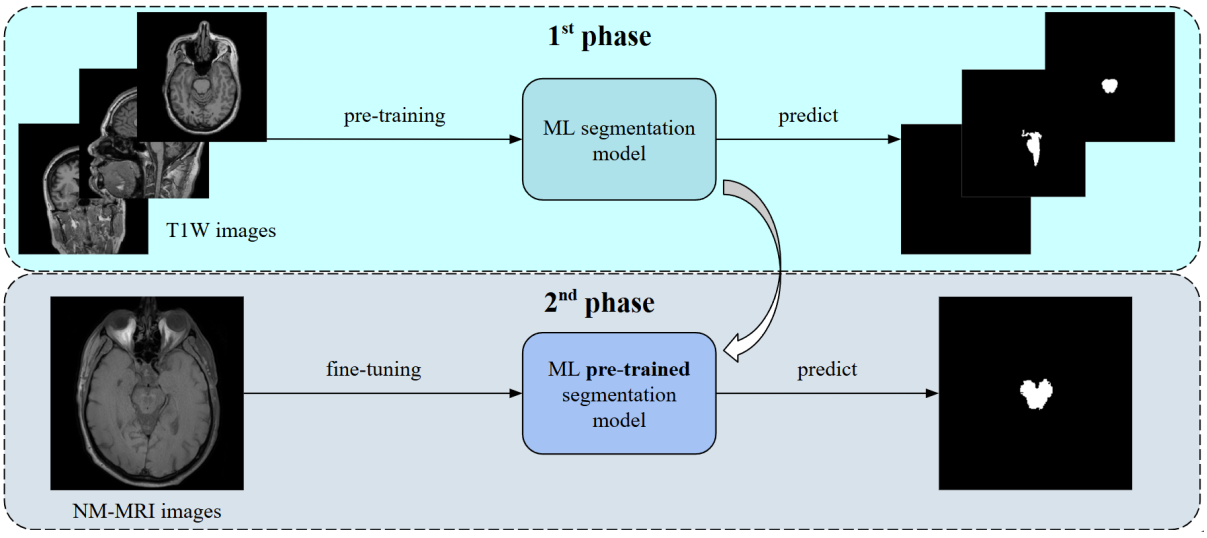


Figure 1.4. T1W axial image of the brain to the left and an NM scan to the right, both from the same patient. The brainstem's borders are highlighted in red in both images, and in the NM example, the NM signal is indicated by the yellow arrow, evidencing its contrast when compared to the T1W case.

Considering that a systematic evaluation of the models trained with T1W scans is the basis to a fine-tuned segmentation model of NM scans afterwards, our work focused on this first step, once if it malfunctions with T1W images, it will fail in segmenting NM images as well. Still, we aspire to make a preliminary case study to provide first impressions on the fine-tuned model and check its potential to succeed in this task, as described in Frame 1.1.

Frame 1.1. Pipeline of the scientific propose considering both phases of the study: the first one that includes the pre-training process of the work and the second one, regarding the fine tuning of the model The inputs and outputs expected from both phases are also evidenced.



In this sense, we aim to assist radiologists and neurologists in the evaluation of a

disorder with such a challenging diagnosis and staging, by coupling T1W MRI scans to well established computer vision models and answering the research question: what's the difference in terms of performance of computer vision models, in this case U-Net adapted architectures, in the task of segmenting the brainstem out of different anatomical planes of T1W neuroimages of PD patients and Control Group (CG), using 2 different sized selected ranges of centralized slices and ground truth masks created by a specialist with the assistance of a piece of software?

1.4 Objectives

1.4.1 General Objective

Our general goal is to evaluate comparatively the ability of optimized U-Net models trained with T1W MRI scans from different anatomical planes, including one trained with all of them, to segment the brainstem in 2 different selected ranges of images obtained from PD and CG patients, with ground truth masks generated by a software and validated by a specialist, using 2 main metrics: Dice Similarity Coefficient (DSC) and Intersection over Union (IoU) and 3 other supporting distance measures: Hausdorff Distance (HD), Symmetrised Modified HD (SMHD), and 95th Percentile HD (HD95). Our focus is on the model trained with axial slices, because following that, we aim to make a first analysis of the segmentation capacity of the axial model, trained with the larger region ranged subset and fine-tuned with NM scans, to detect the brainstem area in NM images.

1.4.2 Specific Objectives

To be able to achieve the general objective, we can consider the following granular accomplishments:

- Define centralized ranges of slices to be used as inputs to the models, containing images with the brainstem, but also images without it;
- Select the optimal hyperparameters of each U-Net model trained with specific types of slices through a grid search;
- Train and evaluate the models' performances using DSC, IoU, HD, SMHD, and HD95 to compare themselves;
- Analyze the statistical significance of the results, checking whether the generic model overcomes the single-plane trained models or not;

- Verify the axial model's performance and potential to extract the brainstem out of T1W images;
- Analyze the performance of a fine-tuned axial model with NM images to verify its potential to detect the structure of brainstem region in these type of scans.

2 Theoretical Foundation and State-of-the-Art of Parkinson's Disease Diagnosis and Staging and Computer Vision in Neurodegenerative Disorders

2.1 Parkinson's Diagnosis and Staging

Although PD was first described more than 200 years ago by James Parkinson [40], in 1817, most of the groundbreaking advances have been done only in the last 50 years [39]. Even so, the observations reported by dr. Parkinson in his essay already contained key elements, as motors and non-motors, that today are understood as cardinal progression signs of PD [40]. For instance, his description of tremors already highlights the difference between resting and action tremors and indicates that in his cases the tremors were mostly at rest. Another 2 important considerations were the unilateral onset alongside with the involvement of hands and arms before legs [39, 40].

Nowadays, it is known that PD's motor symptoms, such as bradykinesia, postural instability and tremors are associated with the loss of dopamine-producing neurons [52]. These dopaminergic neurons are part of a cells group called catecholaminergic nerve cells, which produce neurotransmitters like dopamine, adrenaline and noradrenaline, that among other functions, play an important role in motor control [52]. The dopaminergic cells located in the Substantia Nigra (SN) and the noradrenergic cells located in the Locus Coeruleus (LC) are characteristically pigmented with a substance called NM [20, 56]. In normal aging, the concentration of this molecule tends to increase in these 2 regions over the years, however, inPDpatients, the death of SN and LC neurons leads to the consequential decrease of this pigment, as shown in Figure 2.1. By the time of the diagnosis, 50% or more of these dopamine-producing cells are already dead [20, 34, 56, 65].

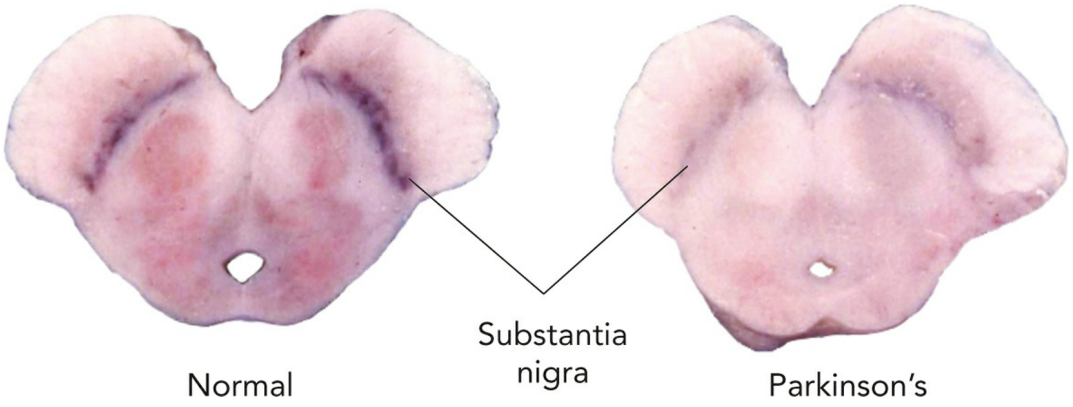


Figure 2.1. To the left, a midbrain slice showing the SN of a control group person and to the right, of someone with PD, highlighting the difference between each of them in terms of dopaminergic cells, seen as dark patches in the SN. Source: [25].

This degeneration is related to the deposition of misfolded alpha-synuclein proteins in the form of Lewy Bodies (LB) in the brainstem and cortical regions [34]. LB are clumps of proteins found by Frederick Lewy, in 1912, inside some neurons in brains of people whose deaths were caused by PD [25]. These malformed proteins, also called prions, are generated from abnormal precursor proteins, probably mutated or damaged, that misfold into a toxic shape, causing the neuron to malfunction and eventually die [25]. An even more aggravating factor is that these prions are capable of self-propagating and inducing healthy precursor proteins to fold abnormally [25]. This process is illustrated in Figure 2.2. Furthermore, prions were found to be related to many other neurodegenerative disorders besides PD, such as Alzheimer's Disease (AD), Post-Traumatic Stress Disorder (PTSD), genetic form of Amyotrofic Lateral Sclerosis (ALS), Huntington's disease, as presented in Figure 2.3.

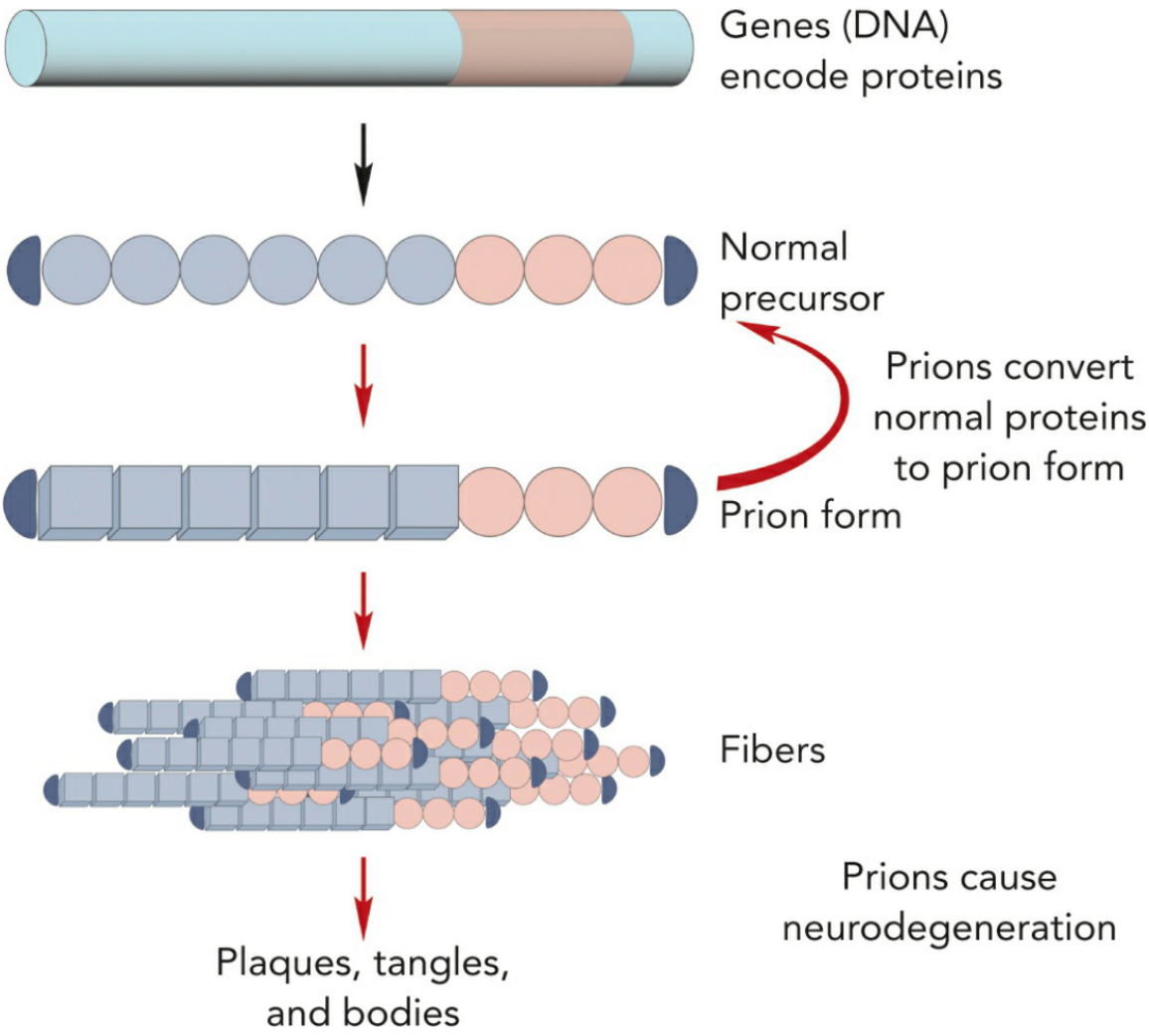


Figure 2.2. Formation of toxic clumps of misfolded proteins that lead to neurodegenerative disorders like PD and AD and induction process of normal precursor proteins into prion form. Source: [25].

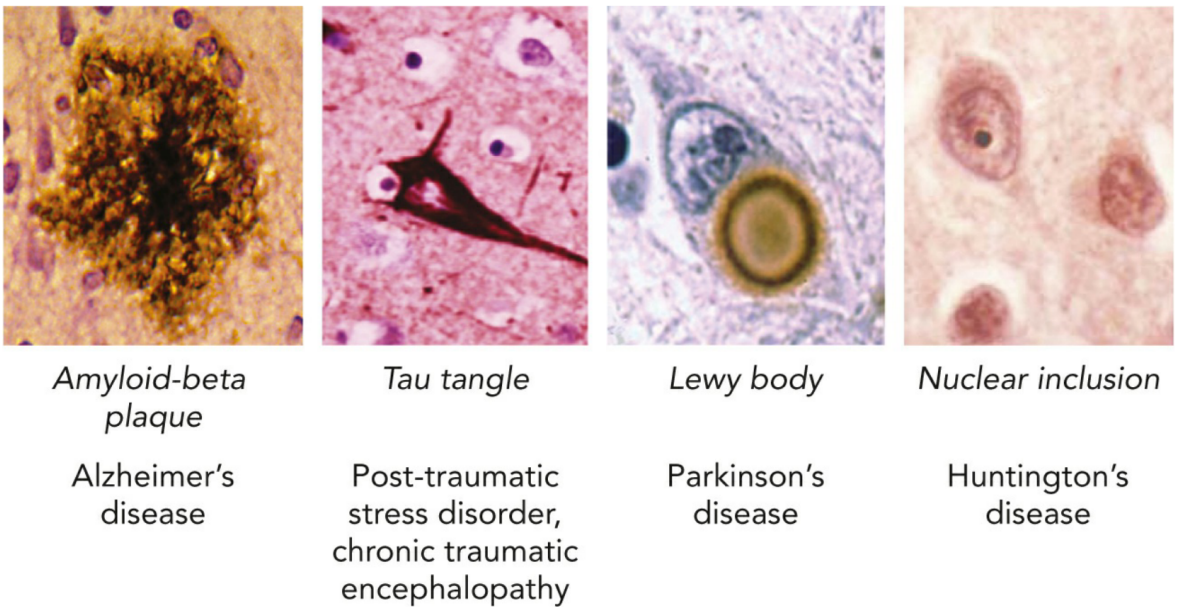


Figure 2.3. Aggregates of misfolded proteins that lead to different neurodegenerative diseases. To the left, amyloid-beta plaque, related to AD. Right next to it, tau tangle, associated with PTSD and Chronic Traumatic Encephalopathy (CTE) (presented by people who had several repeated concussions). In the third position, alpha-synuclein body, also called LB, linked to PD. And, finally, Nuclear Inclusion, found in Huntington's disease patients. Source: [25].

To better understand the progression of PD, in 1967, Margareth Hoehn and Melvin Yahr [21] developed a clinical staging scale for the course of the disease, called after their names. Until nowadays, it is used as a clinical instrument to measure the progression and disability of the disease, classifying it into 5 possible stages, from I to V, with the starting point being the clinical diagnosis [7, 21].

Then, in 2003, Heiko Braak [6] formulated one of today's the most scientifically accepted theory of the disease's onset, that defines anatomopathological stages of PD based on affected structures in the Central Nervous System (CNS) [7]. According to Braak's theory, the first phase starts 20 years prior the clinical diagnosis of the disease, where patients present prodromal symptoms such as Rapid Eye Movement (REM) sleep Behavior Disorder (RBD), depression, hyposmia and constipation [6, 7]. These manifestations of PD can be explained by the 6 anatomic pathological stages of the disease presented by Braak. The first stage involves the autonomic and olfactory centers, which may justify the less frequent bowel movements and the loss of smell, respectively. In the second stage, LB reach the LC, fact that potentially points out to the RBD and depression. Then, in the third stage, the SN is affected, leading to the onset of motor symptoms [6, 43]. From the fourth stage forward, LB attack to non-brainstem structures, until they spread to the cerebral cortex, in stages 5 and 6. A chronological comparison between Hoehn and Yahr's and Braak's scale and the anatomopathological pathway of the disease are detailed

by [57] in Figure 2.4.

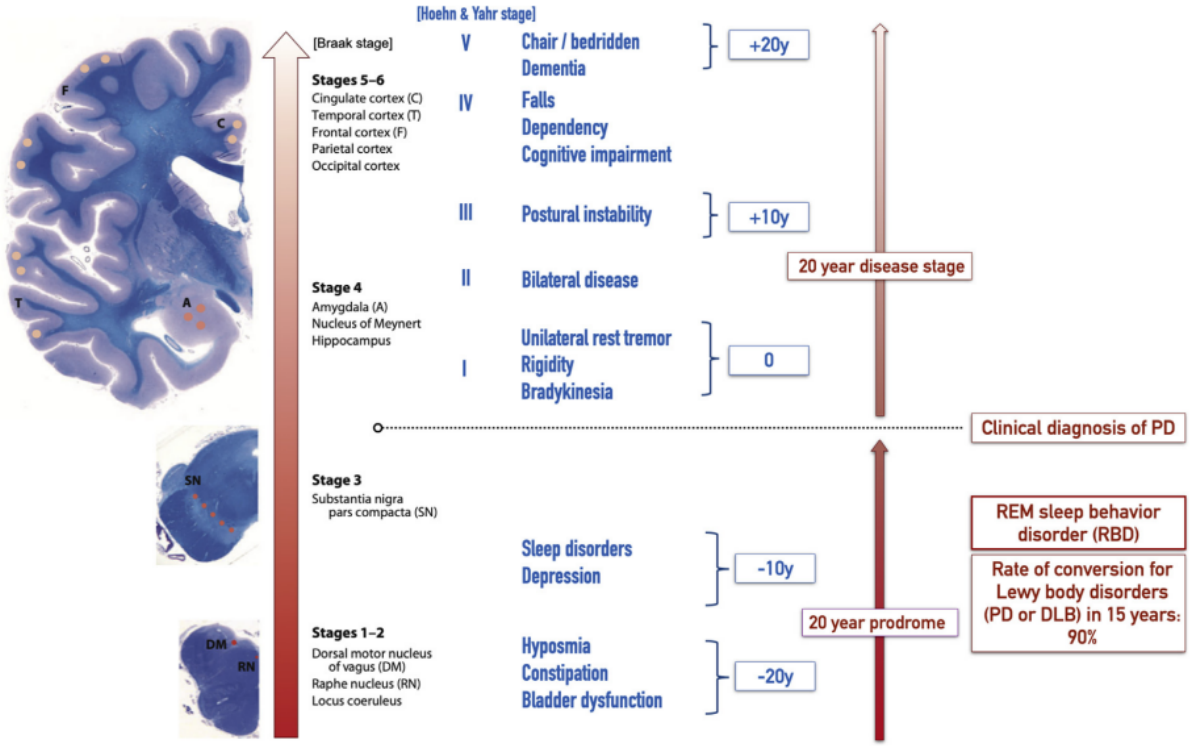


Figure 2.4. Chronological comparison between clinical (Hoehn and Yahr) and anatomopathological (Braak) stages. The arrow to the left in red degrade represents Braak's scale, while the Roman numerals in blue identify Hoehn and Yahr's scale. Also in blue, the symptoms related to each stage of Hoehn and Yahr's scale are described alongside with the time span after the diagnosis of PD [57, 7]. A — amygdala; T — temporal lobe; C — cingulate cortex; SN — SN; F — frontal cortex. Source: [57].

Moreover, the current process of diagnosing PD remains strictly clinical, meaning that doctors will determine the onset of the disease based on a checklist of symptoms, family, medical and medication histories, and physical and neurological exams, as long as there's no conclusive laboratory or imaging test for it [41]. The actual confirmation of the diagnosis can only be done in a postmortem examination [9]. Therefore, making a diagnosis of PD can be very costly, regarding the highly specialized professionals required and the time-consuming medical evaluation [54]. Also, due to its multifactorial and subjective aspects, the chances of false negatives to patients with very early PD and false positives to patients with secondary parkinsonism are high [20].

However, in 2006, Makoto Sasaki et al. [47] proposed an MRI acquisition protocol to obtain NM sensitive MRI scans. Bearing in mind the impossibility to access and monitor the presence and advance of LB in living brains, this method provided a non invasive way to not only make possible the detection of PD in earlier stages, but also pave the way to differentiate other types of Parkinsonian syndromes from PD based on imaging [20, 54]. Many studies have proven a direct raise in the diagnostic accuracy with the support of

the visualization of the signal intensity in SN [1, 9, 20, 27, 30, 43, 39, 48, 54].

Regarding PD's treatment, currently the gold standard for symptomatic PD is still Levodopa, a precursor of dopamine [41]. Despite its notorious effects on motor symptoms, it does not impede the progression of the disease neither treat the non-motor symptoms [34, 35]. Yet, new studies on disease-modifying therapies are increasing, such as prasinezumab [62], inhinitors of LRRK2 [26] and gene therapy [60]. Thus, an early diagnosis allows a wider range of therapy alternatives, leading to more accurate treatment matching, and, at the very least, a delayed progression of the disease.

In light of that, one challenge to an early diagnosis is that the pre-motor symptoms are very heterogeneous and overlap with many other diseases. There are, however, supporting tests, such as genetic testing, blood testing, Computerized Tomography (CT) and MRI, that, along with advanced studies on prodromal symptoms, are preparing the ground for early diagnosis of PD.

2.2 Computer Vision

2.2.1 Machine Learning

According to Mitchell [18, 36] "A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E". So, regarding the types of experiences E, most of ML algorithms can be classified into 2 different categories according to their learning process: unsupervised and supervised [18]. Unsupervised learning programs learn properties and characteristics just from experiencing the dataset [18]. For instance, if you train a model with many images of dogs, once you input a cat image, it will categorize it as a different class than the one it was trained on. In this case, we would not need to provide information on what a dog or a cat is. Oppositely, supervised learning algorithms learn by experiencing a labeled dataset, which means that for each data sample, there is a label or target associated [18]. There are still other categories to the learning process of learning algorithms, such as semi-supervised learning, that couples both previously mentioned methods by using labeled data together with unlabeled data, avoiding generally extensive annotation labor that are generally costly, and reinforcement learning, that learns through the interaction with the environment, similarly to how babies learn how to avoid hazard [18, 66].

A higher-level technique used for systems to learn is transfer learning. It refers to the scenario where a model learns in one domain, utilizing one of the above mentioned methods (supervised learning, for example), and then it is explored in another domain in terms of generalization improvement [18]. In the context of computer vision, an example

would be an algorithm that learned from a great amount of human photos and then, this same model is trained with fewer images of dogs, for instance. In spite of the differences between human and dog, many low-level representations used by the algorithm are shared in these 2 tasks, such as edge detection, lighting changes, and others, which justifies why the second training process requires less samples than the first one to learn the task [18].

In respect to the tasks T, there are 2 major categories of tasks: regression or classification problems [24]. In one hand, when the expected output is a quantitative variable, generally the model's task is a regression. On the other hand, when the response expected is has a categorical or qualitative characteristic, the model's tasks is a classification [24]. In between these 2, there are other tasks such as transcription, translation, anomaly detection, and others [18]. Finally, in order to measure quantitatively the model's performance in these tasks, a measurement P must be chosen. Some of them are accuracy, error rate, and sensitivity, mostly used in classification problems, and also Mean Absolute Error (MAE), Mean Squared Error (MSE), and Mean Absolute Percentage Error (MAPE), usually associated with regression tasks [18].

Now, the most fundamental machine learning model is the Multilayer Perceptron (MLP), which is a feedforward neural network basically composed by an input layer, hidden layers of neurons, and an output layer [18]. These layers define the concept of depth of a model, which leads to the name "deep learning" [29]. Another type of deep learning model is a Convolutional Neural Network (CNN), which is another kind of neural network based on the mathematical operation of convolution [18]. Differently from MLPs whose learning process is generally many multiplications between input and parameter matrices, CNNs are characterized by sparse connectivity, which means that parameters do not necessarily interact with every input [18]. To exemplify how sparse weights influence in a model's efficiency, if we consider an input image with thousands of pixels, kernels with tens of parameters would already be able to detect relevant features, such as edges [18]. As detailed by Goodfellow et al. [18], in terms of time complexity, if m is the number of inputs and n the number of outputs, then we have $m \times n$ parameters and $O(m \times n)$ of runtime in only one sample. Now, if n is reduced to k, a substantially smaller value, then the algorithm's runtime is diminished to $O(m \times k)$, which brings advantages like statistical efficiency and lower memory requirement [18].

2.2.2 Computer Vision in Neurodegenerative Disorders

Meanwhile, despite the introduction of CNN decades ago, in 1989 by Yann LeCun, neural networks applied to computer vision only started to achieve notorious accomplishments in the 2010s [29]. To be more specific, the turning point happened in 2012, with the publication of AlexNet paper, by Alex Krizhevsky et al. Its main novelty was the combi-

nation of the use of Graphic Processing Unit (GPU), which made deeper CNNs training possible, and the application of Rectified Linear Unit (ReLU), a non-saturating activation function that, coupled with regularization techniques enabled a faster convergence of the model [29].

Among the many possible application of computer vision methods, a successful one is in healthcare domains, especially in medical imaging [29]. In addition to the previous mentioned advances in MRI techniques and acquisition protocols, MRI scans' resolution is also enhancing be it because of the higher magnetic fields or because of more sophisticated methods of image reconstruction and treatment. Therefore, a particularly privileged area is neuroimaging, regarding the tiny and irregular-shaped structures of the brain [29].

Now, gathering the discoveries of new possible imaging biomarkers of PD's and the evolution of computer vision methods based on Deep Learning (DL) models, many works are pursuing the goal to segment brain structures in order to support medical decision making, mainly about diagnosis. For instance, Li et al. [31] suggested a 3D region and U-Net-based CNN for segmenting substructures of 3 subcortical neural formations: brainstem, striatum and ventricular system. They used both gold, with manual annotations, and silver, with atlas-based automatic segmentation, standards to create the ground truth masks out of T1W images, with the first labeling method composing only the test set. As metrics, they used DSC, HD95, and Average Symmetric Surface Distance (ASSD). As results, the region-based method was proven to surpass the patch-based approach, and Freesurfer's performance as well. Another result of this work was the application of their proposed model to segment and differentiate, in volumetric terms, patients with different Parkinsonian syndromes and other neurodegenerative disorders, which indicated the midbrain and the pons as crucial.

Sander et al. [46] compared performances of Multi-Dimensional Gated Recurrent Units (MD-GRU) and Freesurfer on segmenting brainstem substructures to manual annotations made by a trained neurologist. The dataset used contained participants diagnosed with Multiple Sclerosis (MS) or Clinically Isolated Syndrome (CIS), AD and CG. Attested by DSC, MD-GRU surpassed Freesurfer's results, when both compared to the gold standard ground truth.

In the same line, Nigro et al. [38] also aimed to segment the midbrain, pons, middle and superior cerebellar peduncles, third ventricle, and frontal horns to differentiate PD patients from Progressive Supranuclear Palsy (PSP) patients and from CG. The fully automated DL method contained a Residual Network with 50 layers (ResNet-50) as encoder and a U-Net as decoder and it was compared to gold standard ground truth masks. The segmentation was measured by DSC and the classification, by Area Under the Receiver Operating Characteristic Curve (AUC).

A binary classification approach was proposed by Wang et al. [58], preceded by SN segmentation. So, the pipeline was to train a Comprehensive Attention-based CNN (CA-Net) coupled with a U-Net with Quantitative Susceptibility Mapping (QSM) and T1W images to segment 5 brain nuclei regions, with SN being one of them. DSC was used to measure its performance. Then, a second model called Attention Gated Squeeze-and-Excitation Residual Networks (suggesting next dimension) with 50 layers (AG-SE-ResNeXt50) was in charge to do the classification, between CG or PD.

Another study, conducted by Dünnwald et al. [12] suggested 2 3D-U-Net-based methods called CoRe-U-Net and Multi-Scale CoRe-U-Net (MS-CoRe-U-Net). Their task was to localize the LC in T1W scans. The ground truth masks were generated with the intersection of 2 manual annotations made by 2 different experts or switching between one another. Particularly, the reference masks of the pons were created by Freesurfer. This study highlighted the efficiency of their model to infer, taking seconds to do the prediction.

Further, Jafari et al. [23] trained a Feedback Weighted U-Net (FU-Net), a modified U-Net, to segment the midbrain and the SN with 3 to 4 axial slices from each of the 102 patients. The results were compared to 2 models, including U-Net itself. In the midbrain segmentation task, this model did not show statistical significance. However, in the SN segmentation, it outperformed, especially when trained with smaller training sets.

Similarly, Le Berre et al. [30] developed 2 adapted U-Net architectures to segment both the midbrain and the SN trained with NM-MRI scans. The final comparison was made with gold standard manual annotations in terms of DSC. An insightful consideration in this study was to use a valid thresholding method to determine whether the pixels were part of the hyperintense areas or not. It calculates the average background signal intensity and sums it to 1.5 times the standard deviation. Their results showed that hyperinstense areas of NM are significantly reduced in PD patients when compared to CG, even with the low precision segmentation of the SN.

Lastly, NigraNet was suggested by Gaurav et al. [16], a U-Net-Based model designed to assess nigral NM. Trained with NM-MRI images, this model was also evaluated under DSC metric and compared to manual annotations. Their main achievements were to use a training set of almost half the size of the test set, highlighting the model's generalization capacity under small dataset contexts, and the detection of PD in patients in earlier stages than the ones identified in other works, such as in the work of Le Berre et al. [30].

2.2.3 Performance Metrics for Medical Image Segmentation Tasks

Regarding the above mentioned related works, there are some particular metrics that appeared more frequently. Some of them are DSC, also known as F-measure or F1-score and IoU, also known as Jaccard index. DSC is a result of an harmonic mean between sensitivity, also called recall or true positive rate, and precision, 2 other metrics arising from the confusion matrix, a two-by-two table constructed after the results of a classifier in terms of its instances, as represented in 2.5 [13]. Precision can be described as

$$precision = \frac{TP}{TP + FP}, \tag{2.1}$$

and sensitivity as

$$sensitivity = \frac{TP}{TP + FN}, \tag{2.2}$$

where TP denotes true positive instances, FP false positive, FN false negative. DSC is an operation between these two

$$DSC = 2 \times \frac{precision \times sensitivity}{precision + sensitivity}, \tag{2.3}$$

that results in

$$DSC = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2.4}$$

or, in terms of set theory,

$$DSC(A, B) = \frac{2 \times |A \cap B|}{|A| + |B|}, \tag{2.5}$$

where A denotes the ground truth region and B, the predicted.

As for IoU, it can also be expressed both in terms of confusion matrix

$$IoU = \frac{TP}{FP + TP + FN}, \tag{2.6}$$

and also in terms of set theory

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|}, \tag{2.7}$$

where TP indicates true positive instances, FP false positive, and FN false negative, just like for DSC. Equally, A represents the ground truth region and B, the predicted. An illustration made by Yeghiazaryan V. and Voiculescu I. [63] shows how confusion matrix

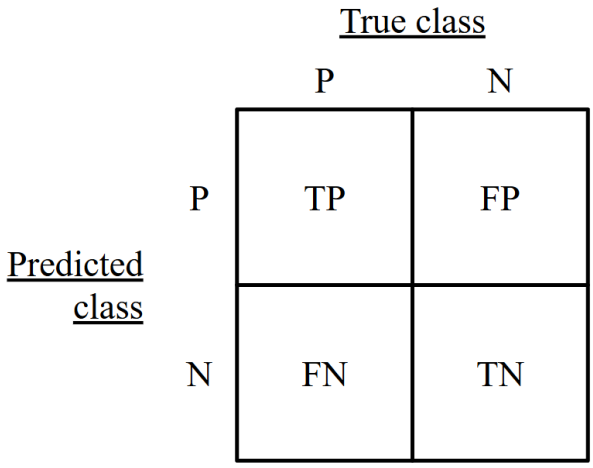


Figure 2.5. Confusion matrix, a disposition of a classifier's results that generates many classification metrics, such as precision, sensitivity and accuracy. P — positive; N — negative; TP — true positive; TN — true negative; FP — false positive;FN — false negative.

relates directly to set theory in Figure 2.6.

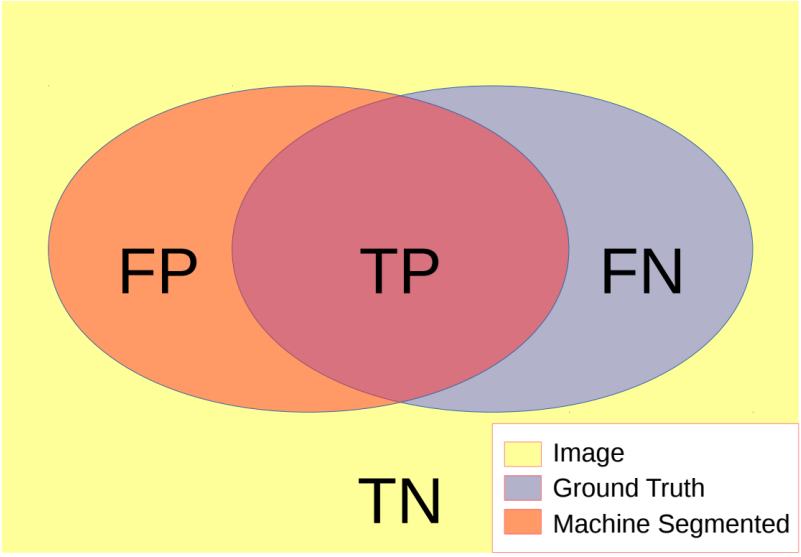


Figure 2.6. Illustration of how confusion matrix terms relate to set theory considering a segmentation task, where the overlapping area refers to the true positive (TP) cases, the remanent area of the ground truth region represents the false negative (FN), the remanent area of the segmented region indicates the false positive (FP), and all the residual area of the image not contemplated by any of the sets refers to the true negative (TN) cases. Source: [63].

According to Müller et al. [37], for medical image segmentation tasks, DSC and IoU are highly recommended due to their focus on true positive classifications and disregard for the true negatives. This recommendation is explained by the imbalanced characteristic of medical images, whose background usually represents the majority of pixels or voxels in a scan when compared to the Region of Interest (ROI) [37]. In other words, if a model

ponders the true negatives as much as true positives, it would still reach high performance metrics even if it predicts the whole image as background.

Some other metrics used to measure performance in other segmentation studies were spatial distance based metrics, such as ASSD, HD and its variations. Differently from the previous explained metrics, these calculate the distance between the ground truth region and the predicted [3]. ASSD, for instance, computes the average of the shortest distances between each point of two surface's boundary, and it is given by

$$ASSD(A,B) = \frac{\sum_{a\in A} \underset{b\in B}{min}(d(a,b)) + \sum_{b\in B} \underset{a\in A}{min}(d(b,a))}{|A| + |B|}, \tag{2.8}$$

where A and B represent ground truth and prediction sets. As for the HD metric, it can be described as the maximum among all distances from a point in one set to the closest point in the other one [31], which means

$$HD(A,B) = max(\underset{a\in A}{max}(\underset{b\in B}{min}(d(a,b))), \underset{b\in B}{max}(\underset{a\in A}{min}(d(a,b)))), \tag{2.9}$$

considering A and B as ground truth and prediction sets. One pitfall of this measure is its vulnerability to outliers. Therefore, alternatively, SMHD is widely used since it minimizes the influence of outliers by calculating the average of the minimum distances between points of one set to the other and vice versa and taking the maximum value among these 2. It is determined by

$$HD_{SM}(A,B) = max(\frac{1}{|A|} \times \sum_{a\in A}(\underset{b\in B}{min}(d(a,b))), \frac{1}{|B|} \times \sum_{b\in B}(\underset{a\in A}{min}(d(a,b)))), \tag{2.10}$$

where A and B represent the ground truth and the prediction sets. Meanwhile, HD95 is defined by the same procedure as the standard HD, except that the highest 5% of the resulting maximum distances of both sets are excluded [31].

3 Materials and Methods

In this chapter, we present the dataset used for the models training and testing, describing how it was acquired, analyzed, processed, and augmented. Later, we explain U-Net, the model we chose to study and apply to our database, detailing each layer within its structure. Next, we introduce how the hyperparameter optimization was done and, lastly, which metrics we used to measure the models' performances.

3.1 Dataset and Image Preprocessing

3.1.1 Dataset Acquisition

The neuroimages that compose the dataset were acquired initially for another study on cognitive impairment of patients with PD back in 2021 [7]. The machine used to obtain the scans was a Philips Achieva 3.0 Tesla, equipped with a coil SENSE of 8 channels, located in Santa Marta Hospital (Taguatinga-DF). Also, the neuroradiologist made the acquisition similarly to the Alzheimer's Disease Neuroimaging Initiative-3 (ADNI-3) study, as described in detail in [7].

The T1W scans were 3D acquisitions 256x256x256 sized, while the NM images were 2D scans with size of 584x584. The NM acquisitions were done perpendicularly to the biggest axis of the brainstem and were limited to the scans where the SN was visible, resulting in 12 slices. In order to be able to use the same model with both datasets, we reduced the dimensions of the NM scans to the same as T1W, which is explained in the next sections. In Figure 3.1 we present an example of both types of scans from the same patient in an approximate slice, while Table 3.1 presents summarized information on level of education, average age and gender distribution between the T1W and NM scans.

Ground truth masks from the T1W scans were obtained with the assistance of Freesurfer piece of software [14]. Freesurfer's automatic subcortical segmentation was based on an atlas containing probabilistic information on the location of 37 brain structures, that assigns each voxel to one of these 37 classes [15]. One of these labels is called "Ventral Diencephalon (DC)", a term that refers to a group of small structures that cannot be distinguished in standard MRI acquisitions [4, 49]. Among them, there is

Table 3.1. Information on participants of the study regarding the type of MRI acquisition protocol, average age, average education (both measured in years), and the proportion between males and females.

	Average age (years)	Average education (years)	Males	Females
T1W scans*	60.69	15.21	40	21
NM scans*	60.94	15.71	35	14

* information on 3 subjects that participated in the study were missing, which justifies the total number of participants differentiating from the reported in 3.2.

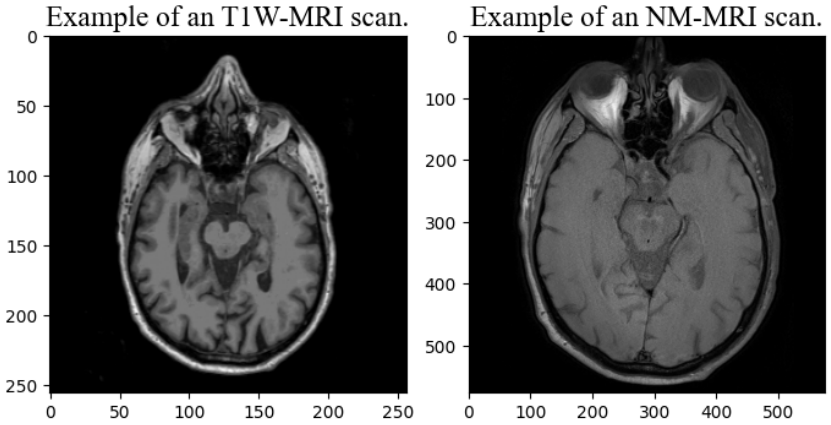


Figure 3.1. Example of an T1W scan in contrast to an NM image. The differences do not only refer to their sizes or acquisition angle, but also to the contrasting signals evidenced.

the ventral tegmentum, that is actually part of the midbrain and, therefore, the brainstem [32]. In light of that, to contemplate the whole region of the brainstem, specially the midbrain, we selected 3 labels to compose our brainstem's ground truth masks: brainstem, left-ventral DC and right-ventral DC, as illustrated in Figure 3.2. Additionally, Freesurfer's outputs were meticulously analyzed by a neuroradiologist and a neurologist in a previous study [7].

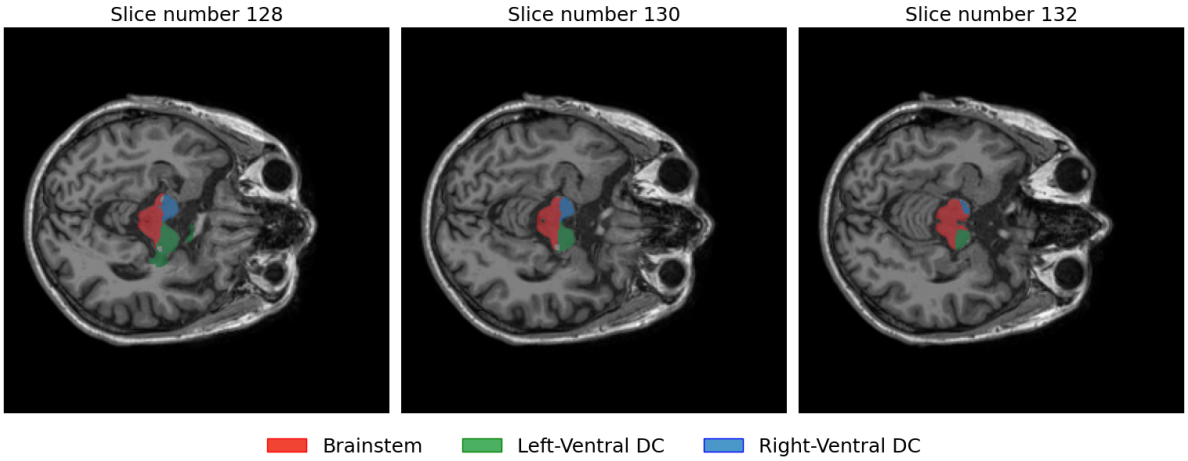


Figure 3.2. Example of 3 axial slices of a T1W MRI scan with its Freesurfer's masks identified, contemplating the brainstem, and the left and right ventral of the DC.

3.1.2 Image Selection and Preprocessing

Firstly, a total of 13 patients' images were excluded from the dataset due to clinical criteria and quality issues related to the masking and imaging process. It resulted in a sample of 64 patients: 51 with PD and 13 from CG, as shown in Figure 3.2. Among these 13 excluded cases, 9 were due to clinical criteria, such as patients with other types of parkinsonism, 3 were due to encephalomalacias (softening or loss of brain tissue after an injury) in the scans, and 1 due to segmentation issues.

Then, we chose 2 central ranges of slices from the T1W images to focus on the portion of the brain where the brainstem can be visualized. This culminated into 2 subsets: the small region subset and the large region subset. The first one contemplates the interval of the 122nd slice to the 132nd and the second, the 115th slice to the 165th, as represented in Figure 3.3. This means that in the large region subset, it was considered 50 slices per patient while in the small region subset it was 10 slices per patient.

So, the final composition of the subsets was of 640 images for the small region subset and 3200 images for the large region subset. These subsets were later divided into training and testing sets, where the first was equivalent to 75% and the second to 25% of the total

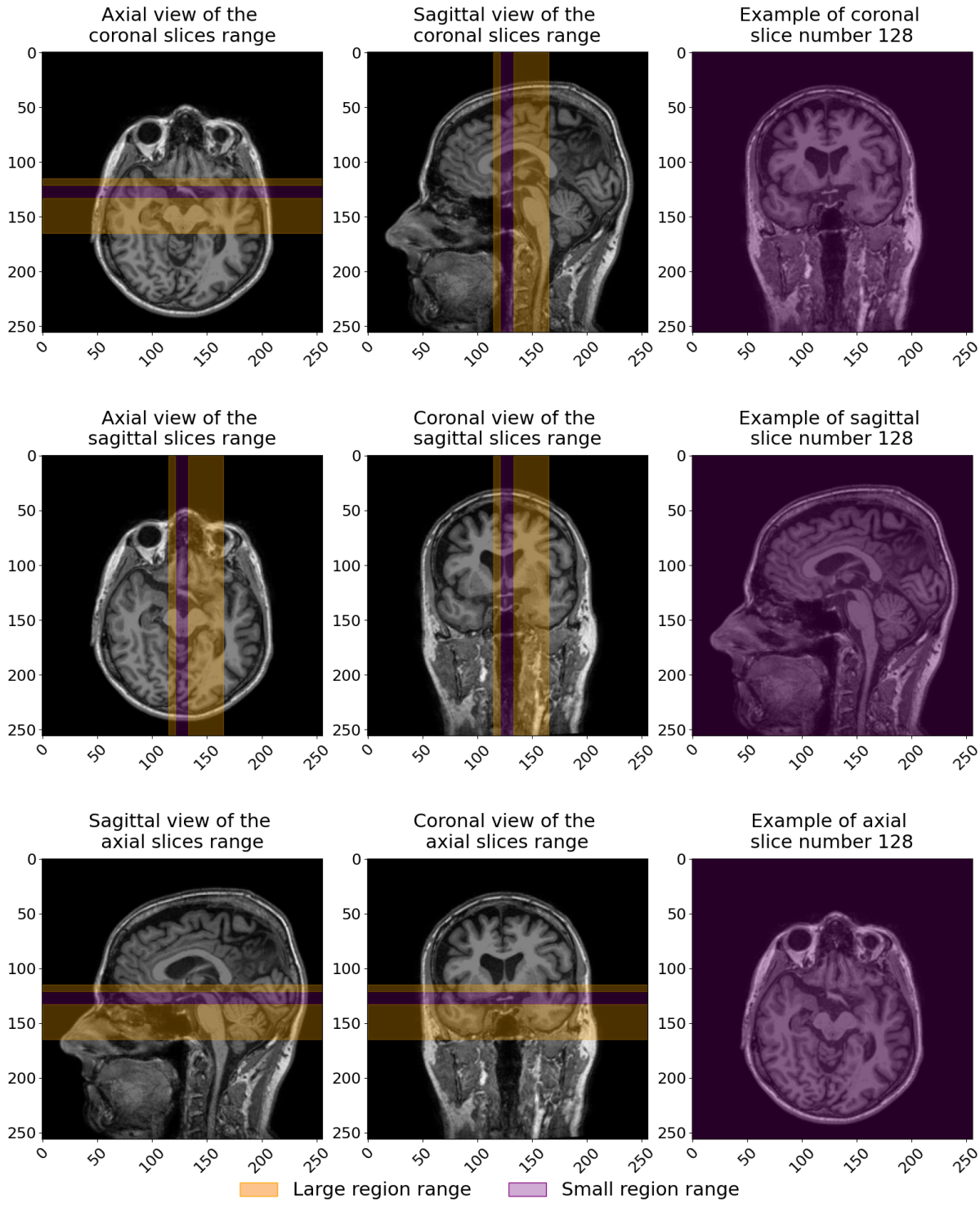


Figure 3.3. Views from the 3 anatomical planes of the slices ranges considered to create the large region subset (115th to 165th slices) and small region subset (122nd to 132nd slices).

Table 3.2. Number of PD patients and CG participants that were included and excluded from the analysis.

		PD	CG	Total
T1W scans	Included	51	13	64
	Excluded	9	4	13
NM scans	Included	41	11	52
	Excluded	8	3	11

number of scans. Also, all images were normalized, ranging from a minimum value of 0 to a maximum value of 1, according to

$$I_n = \frac{I - \min(I)}{\max(I - \min(I))}, \tag{3.1}$$

where I denotes the non normalized image, I_n the normalized image, min(x) the minimum value of all pixels in x, and max(x) the maximum value of all pixels in x.

3.1.3 Data Augmentation

After image selection and training and testing sets division, we applied data augmentation to the training set of both data subsets. The final number of images that composed the training set was 5000. This means that in the large region subset there was 2400 original images and 2600 artificially generated and in the small region subset there was 480 original scans and 4520 resulting from the data augmentation process, as it's elucidated in table 3.3.

Table 3.3. Final division of the data subsets considering training and testing sets after the data augmentation process.

		Original	Augmented	Total
Small region subset	Training	480	4520	5000
	Testing	160	-	160
	Total	640	4520	5160
Large region subset	Training	2400	2600	5000
	Testing	800	-	800
	Total	3200	2600	5800

The operations used in the augmentation procedure were: rotation and noise addition. The range of the rotation angles varied from -30° to 30°, with steps of 10°. Now, the white Gaussian noise had a standard deviation of 0.02 and it was added to every other image, resulting in noise applied to half of the rotated images. This operation of increasing artificially the number of samples in the training set aim to create diverse scenarios for

the model to learn how to generalize the structure of the brainstem. It does not attempt to generate necessarily realistic scans in terms of MRI acquisition protocols.

3.2 U-Net Model for T1W Segmentation of the Brainstem

U-Net is a type of CNN architecture developed in 2015 initially for segmenting neuronal structures in microscopic images, that showed applicability to other biomedical segmentation tasks [29, 45]. It was named after its U shape design, as can be seen in Figure 3.4. The first half of the model is called encoder. It downsamples the input until it reaches a feature map with size 28x28x1024. The second symmetric half is called decoder, that upsamples the feature map until it gets a size of 388x388xK, where K represents the number of classes to segment.

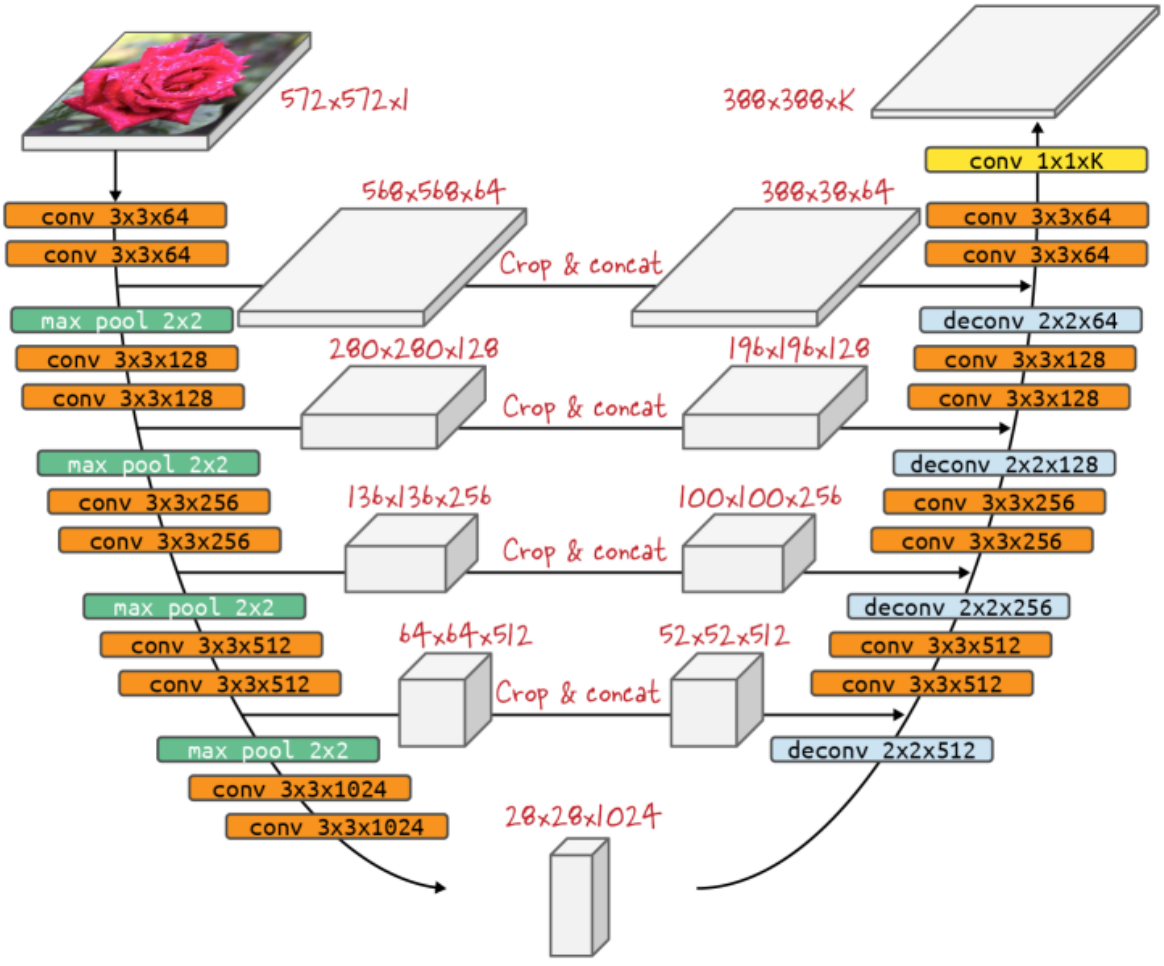


Figure 3.4. Original U-Net architecture, designed by [29], with an input sized 572x272x1 pixels. Each white cube represents a feature map and right on top of it, there is its dimensions. The colored rectangles in the U shape axis denote operations applied to the image. The "crop and concat" skip connections retrieve information from the encoder block to the decoder's inputs along the way. Source: [29].

In terms of architecture, the main difference between the U-Net model and an autoencoder, that is also composed by an encoder and a decoder, is the skip connections that concatenate to the decoder feature maps from the encoder, recapturing relevant spacial information extracted in this block [45]. Now, considering their purposes, an autoencoder task is to learn how to copy its input to its output [18]. Conversely, U-Net was trained specifically to the task of biomedical image segmentation [45]. So, U-Net's output is a segmentation mask and an autoencoder's is the input.

According to the recent work of [52], CNN is one of the methods that achieved more successful results in the context of medical image-based approaches to assist in PD diagnosis. Among 11 MRI-based approaches reviewed, 3 of them used U-Net or U-Net based models, one of them is even called Nigra-Net, refering to SN.

Thereby, we chose U-Net architecture to perform the task of segmenting the brainstem in T1W MRI scans. For that, we made 2 adaptations in the model's structure. First we changed the input size, that in our case was 256x256x1, since the images were in grayscale (with only 1 channel). Second, we considered the number of in the first layer as a hyperparameter to be optimized as well in order to be able to reduce computational resources in situations where the model's performance was equivalent or even better with less filters than U-Net's original number, which is 64. The adapted version of the model used in this study can be seen in Figure 3.5.

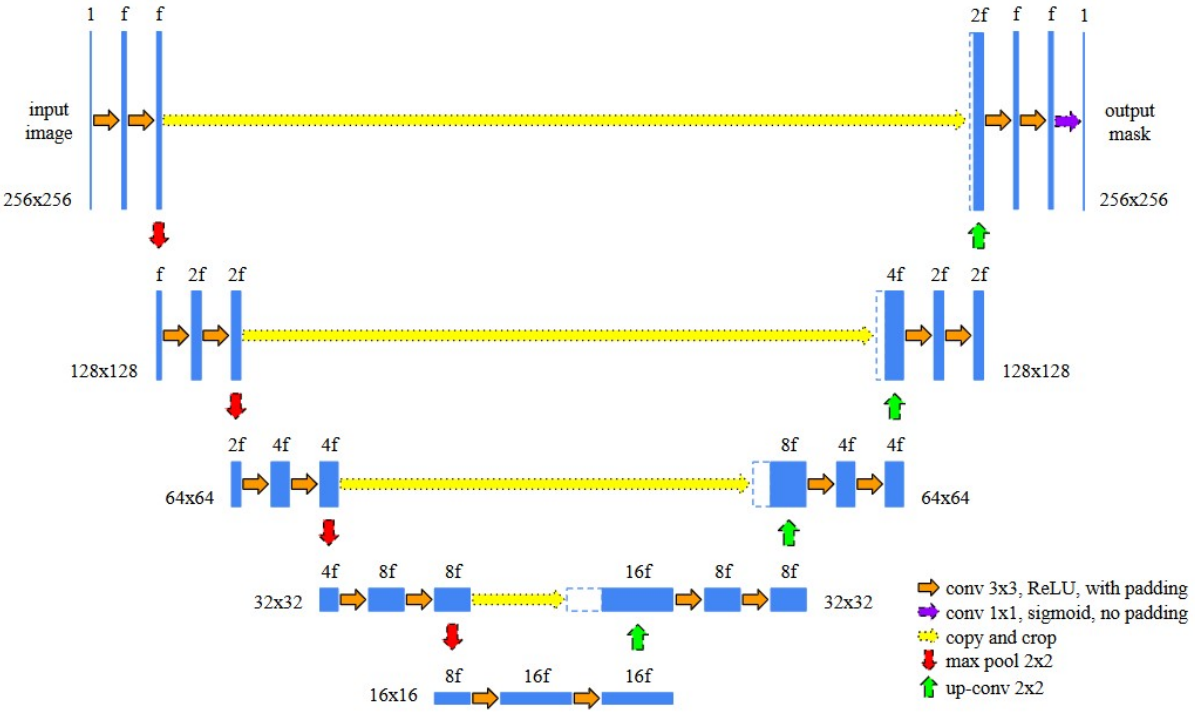


Figure 3.5. The adjusted U-Net with input size of 256x256x1 and number of filters in the first layer indicated as a letter f right on top of each blue box. The arrows denote operations to the feature maps, which are represented by these blue boxes. By their side, there is their dimensions.

3.2.1 Convolutional Neural Networks (CNN)

Most of U-Net layers are CNNs, a category of neural network that differs from other Artificial Neural Network (ANN)s mostly because of its capability to operate in high-dimensional data, such as images and videos [28, 29]. It works by applying convolution operations on the inputs in order to extract local and subsequently, higher-level features [28].

Differently from humans vision system, that can see objects based on the continuous physical process of illumination and reflectance, computers can only "observe" an object through a digital image, which converts a continuous data into a discrete data, composed by pixels associated with numerical values, as shown in the diagram 3.6 [17].

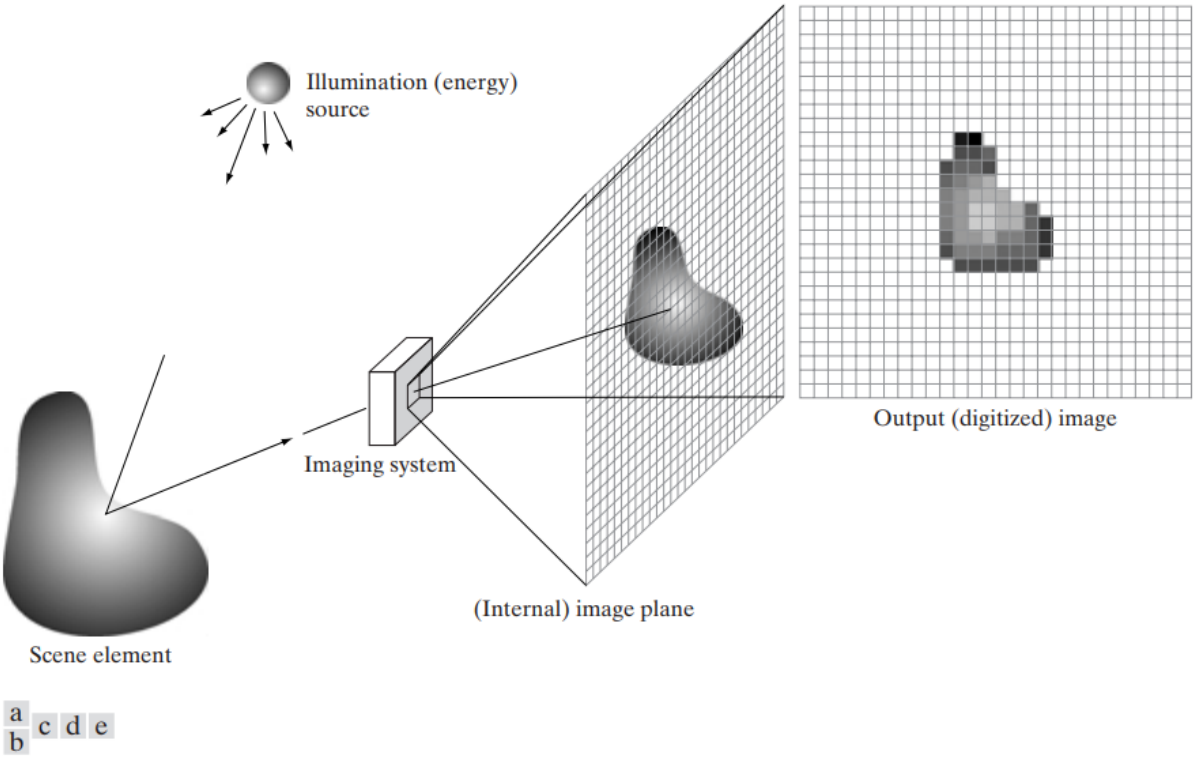


Figure 3.6. An illustration of the image digitization process. (a) Illumination source. (b) An observed object. (c) Imaging system, capturing the amount of illumination reflected by the element. (d) Projection of the scene in an internal image plane, still continuous. (e) Digitized image, product of sampling and quantization processes. Source: [17].

Considering that, what a convolution operation does is apply a filter, that's a usually smaller matrix with discrete numbers, to this image in a sliding way, resulting in another matrix where each element corresponds to the sum of the products between correspondent elements, as it is shown in Figure 3.7 [28]. In CNNs, filter's weights are normally initialized randomly and then, during the training process, they begin to be updated according to the proposed task [28].

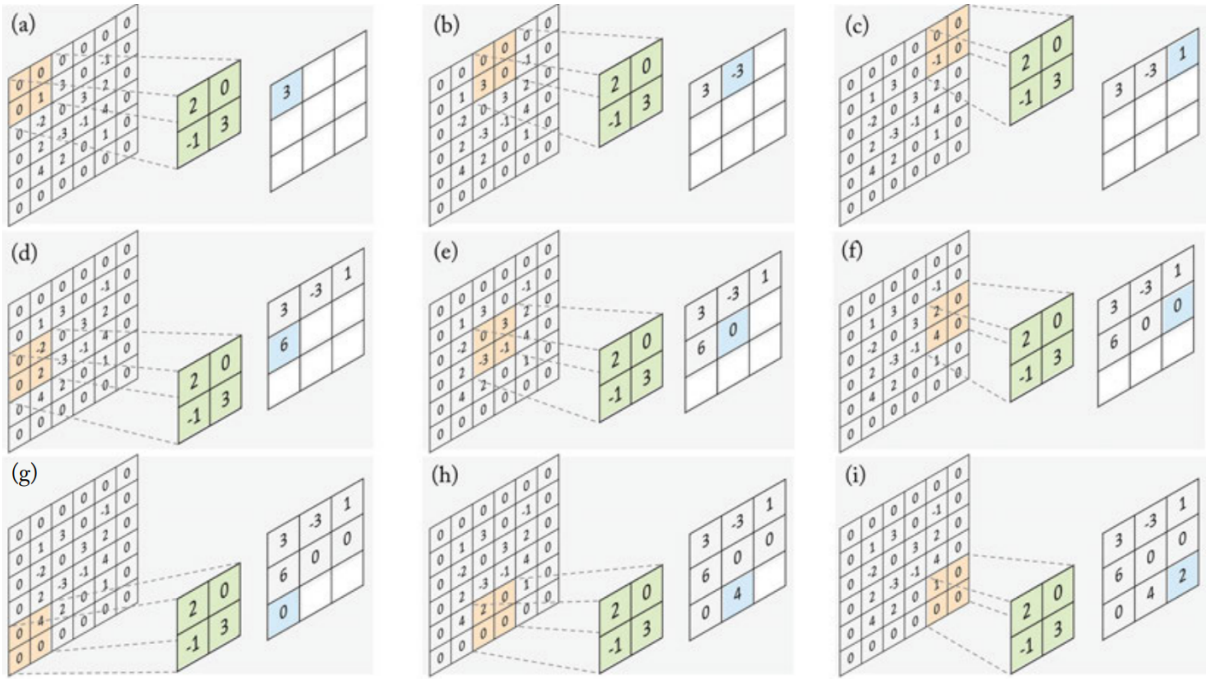


Figure 3.7. An example of a step-by-step convolution between a 4x4 input and a 2x2 filter, with a zero-padding and a stride of 2. The orange area is the one being multiplied by the filter, in green. The result of each step is shown in blue, as being the sum of the products between the orange and the green matrices. Source: [28].

Some important configuration aspects of a convolution are stride and padding. The first refers to the size of the steps given by the convolutional filter when sliding through an input, horizontal and vertically. In the illustration 3.7, a stride of 2 was used. The second term indicates the addition of a border of zero elements around the input image. This technique avoids the reduction of spatial size after the convolution operation. Taking this into account, there are 3 types of convolutions: valid, same and full convolution, with valid convolution having no padding, same having padding to maintain the output's size same as the input's and full having the maximum possible padding [28].

3.2.1.1 Max-Pooling Layers

Another recurrent block that composes a U-Net is max-pooling. This is a type of pooling layer that extracts the maximum value in a determined sliding region of the feature map, ruled by a predefined stride value. In Figure 3.8, max-pooling is defined by 2x2 sliding window with a stride of 1. Another popular type of pooling layer is average-pooling, which, instead of computing the maximum value within a patch, calculates the average of the elements.

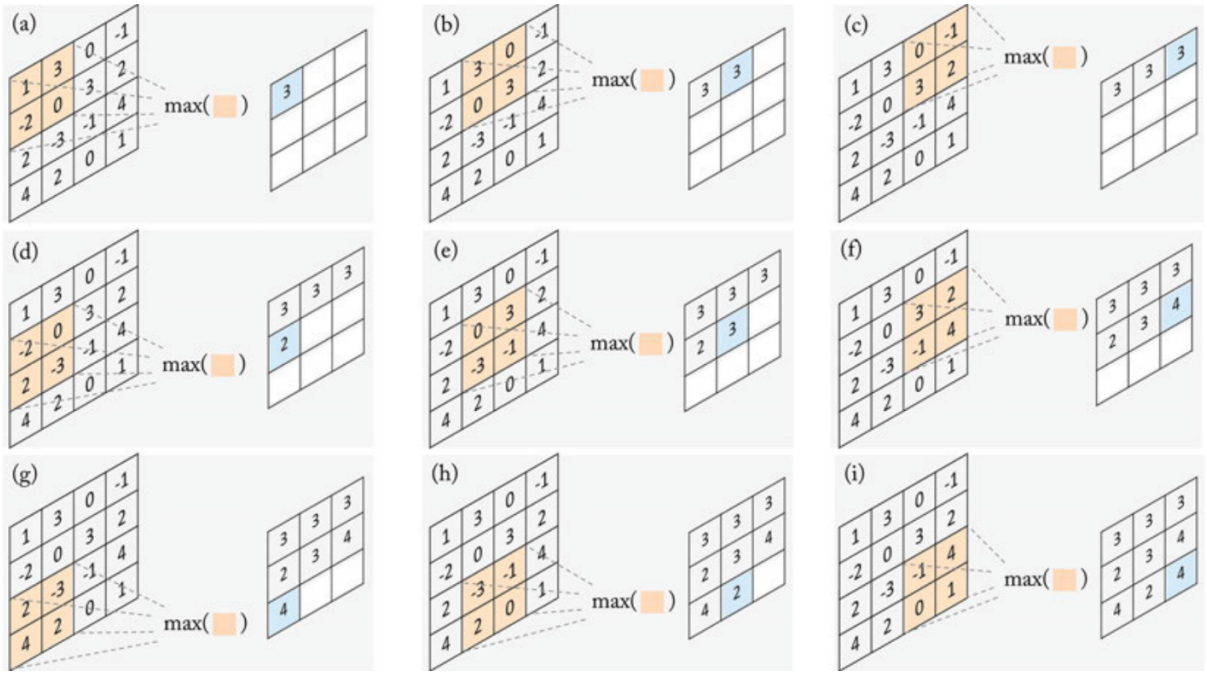


Figure 3.8. A step-by-step max-pooling operation with a pooling region of 2x2, represented by the orange color, and stride of 1. The blue value indicates the resulting maximum value of the pooling area. Source: [28].

3.2.1.2 Fully Connected Layers

Usually, at the end of a CNN there is a fully connected layer which basically flattens the feature maps of the previous layer and feeds it to an activation function subsequently [29], as illustrated in Figure 3.9. Another way to visualize how a fully connected layer works is to consider that it is a convolutional operation with a 1x1 filter that results in the following equation:

$$\mathbf{y} = f(\mathbf{W^T x} + \mathbf{b}), \tag{3.2}$$

where $\mathbf{y}$ and $\mathbf{x}$ represent the output and input vectors, respectively, $\mathbf{W^T}$ the weight matrix of the connections, $\mathbf{b}$ the vector of bias values and f the activation function, typically a softmax in cases of multiclass classification problems [28].

Since our goal using U-Net is a 1-class semantic segmentation, in other words, a binary classification of each pixel of the image, we did not use softmax. Instead, we chose sigmoid activation function, characterized by an S-shape curve ranging from 0 to 1.

3.2.1.3 Transposed Convolution Layers

Finally, the called "up-convolution" layers, which are actually transposed convolutions (also called deconvolutions interchangeably, in the context of computer vision), are

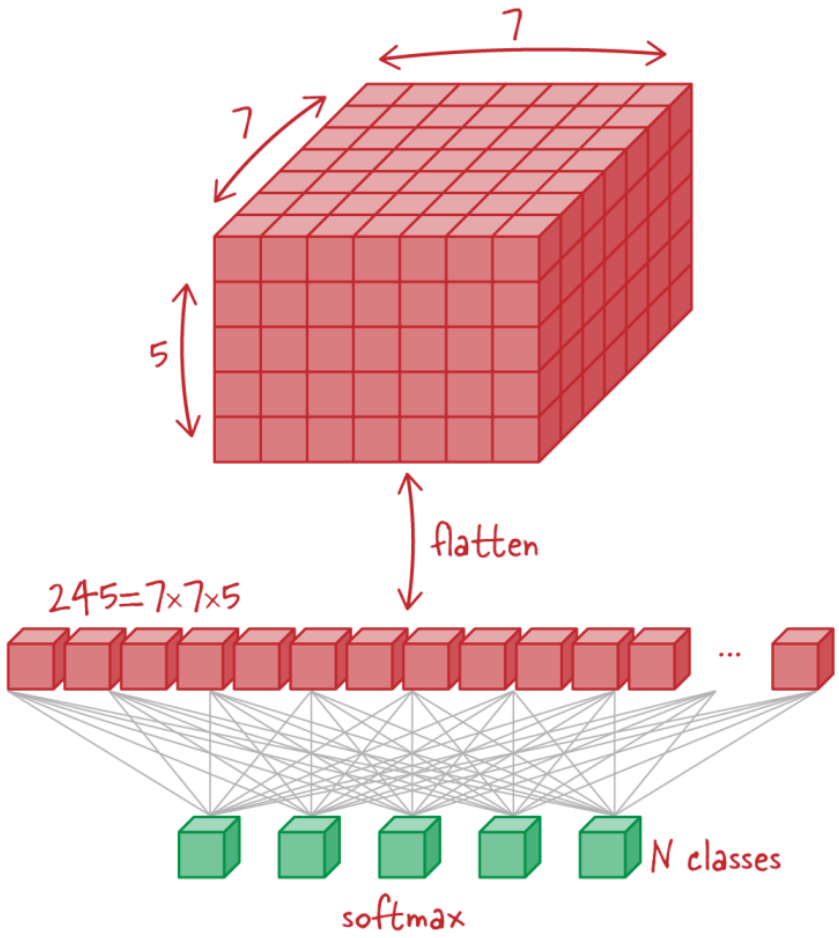


Figure 3.9. An illustration of how a fully connected layer works. A feature map of size 7x7x5 is flattened into a vector with size 1x245. Next, the 1-dimensional data is fed into a softmax activation function, which produces a probability distribution over the N possible classes, ensuring that all probabilities sum to 1. Source: [29].

the protagonists of the U-Net's decoder block. The transposed convolution layers upsample the feature maps inputs with the use of learnable weights. For that to happen, each value in the input multiplies all elements of the filter and the filter sized resultant matrix is added to the output. The following resultant matrices are also added to the output with a specified stride, leading to overlapping values that are summed to each other, generating the upsampled feature map [29]. This whole process is represented in Figure 3.10.

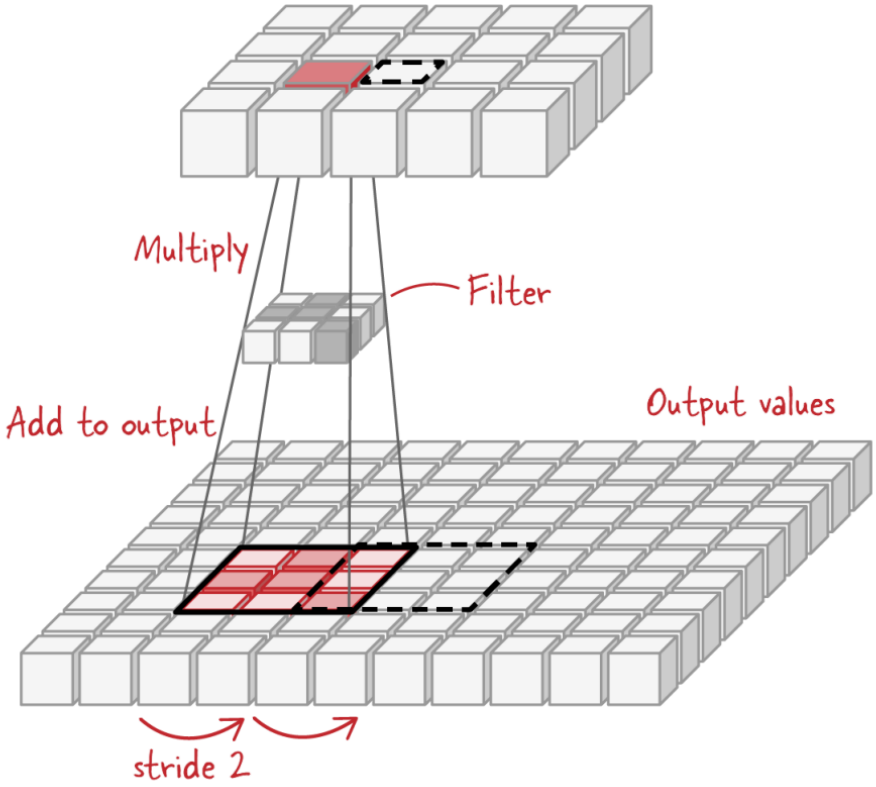


Figure 3.10. A representation of a transposed convolution, where the input is located at the top, the filter right at the center and the output at the bottom. This example uses a stride of 2 and the overlapping elements are evidenced with the intersection of the dotted contour and the continuous contour on the output space. Source: [29].

3.2.2 Hyperparameters and Optimization

We used grid search to find the optimal models, varying 3 main hyperparameters: optimizer, learning rate and number of filters in the first layer. The other 5 hyperparameters we set were fixed, as shown in table 3.4.

The optimizer options were Adaptative Moment Estimation (Adam) and Adaptative Moment Estimation with Decoupled Weight Decay (AdamW), which are usually recommended for computer vision problems [29]. The main difference between these 2 is that

the first adds weight decay to the loss function, while the second, directly to the parameter update step. The effect of this is mostly related to the interference with adaptive learning rates and, consequently, to an optimal convergence. The loss function was set as binary cross-entropy and the activation function as ReLU. Learning rate alternatives were 1e-4 and 5e-4, considering a smaller and a larger step towards convergence. Furthermore, the number of filters in the first layer ranged from 8 to 64, in powers of 2, with 64 being the U-Net's original number of initial filters. Next, epochs were fixed in 300, even though most models hardly ever reached this value due to the early stopping feature, which ends the training after a determined number of epochs without improved performance. This number is called patience, which we defined as 10. Lastly, batch size was kept 1, as used by U-Net's authors [45].

Table 3.4. Hyperparameters range used during grid search. Loss, activation function, number of epochs, patience (maximum number of epochs trained without improvement), and batch size were fixed, while optimizer, learning rate and number of filters in the first layer varied. The optimization was composed of 12 possible combinations per model, once 8, 16 and 32 were the values considered for filters in the first layer of models of singles types of slices, and 16, 32 and 64, for the multi-slice models.

Hyperparameter	Value
Optimizer	[Adam, AdamW]
Loss	Binary cross-entropy
Activation function	ReLU
Learning rate	[1e-4, 5e-4]
Number of filters in the first layer	[8, 16, 32, 64]*
Epochs	300
Patience	10
Batch size	1

* the range of [8, 16, 32] was used in models trained with only one anatomical plane and for the multiplanar model, the range was [16, 32, 64].

Another hyperparameter used, but external to the training process, was threshold, to binarize the outputs. We set a fixed value of 0.5, meaning that pixels with a signal greater than or equal to 0.5 were considered 1 and the others were considered 0. Also, all random seeds were set to 42. However, the use of GPUs for training hinders the exact reproducibility of the experiments [50].

3.2.3 Metrics

Accordingly to the metrics used in related works, we chose 2 main metrics to measure the models' performance, the DSC and the IoU. Regarding the imbalanced characteristic of neuroimages, by choosing these metrics, we avoid being misguided by measures that

are calculated pondering equally true negative and true positive pixels, as mentioned previously in Chapter 2 [37]. Also, we considered other 3 distance based measures for support, which are HD and 2 variations of it: SMHD and HD95.

The first 2 metrics are calculated in terms of percentage, which means that their range in between 0 and 1 and the higher the value, the better the performance. On the other hand, distance metrics are interpret oppositely, ranging from 0 to infinite, the smaller the distance between the ground truth region and the prediction region, the better the segmentation.

3.3 Fine-tuned U-Net Model for NM Segmentation of the Brainstem

This first analysis of the axial model trained with the large region dataset and fine-tuned with NM scans aimed to verify its feasibility and potential. The selection of the model was based on the similarity between T1W scans to NM slices, due to its oblique acquisition, and on the resultant metrics, that showed prevalence of the large region subset trained model over the small region subset trained model. Hence, with the assistance of Freesurfer, our research group created the reference masks by converting the masks generated in T1W space to NM space. Thus, in possession of the scans and the ground truth masks, we were able to fine-tune the selected model. Therefore, we chose the combination of hyperparameters that lead to the best metrics for this model and trained it with these NM images. Once the model's input size was defined based on the smallest scans used to train it, hence the T1W images, we reduced the size of the NM slices previously.

This dimension reduction applied to the NM images was made with a cropping operation, reducing firstly the 576x576 images to the size of 512x512, disregarding 32 pixels of each border of the scans. Then, a decimation downsampling was made by skipping every other pixel, resulting in a 256x256-sized image, as presented in Figure 3.11. This way, we could input this new dataset to fine-tune the model.

Next, we separated the dataset into training and testing sets, with the same proportion used previously (75% for training and 25% for testing) and applied data augmentation with a factor of 10 only to the training set, generating a total of 4212 new images that, summed to the original ones, resulted in 4680 samples. These numbers are shown in Table 3.5. The same parametrization used to augment the T1W images was considered to this set, so rotation and noise mentioned in Subsection 3.1.3 were the operations used.

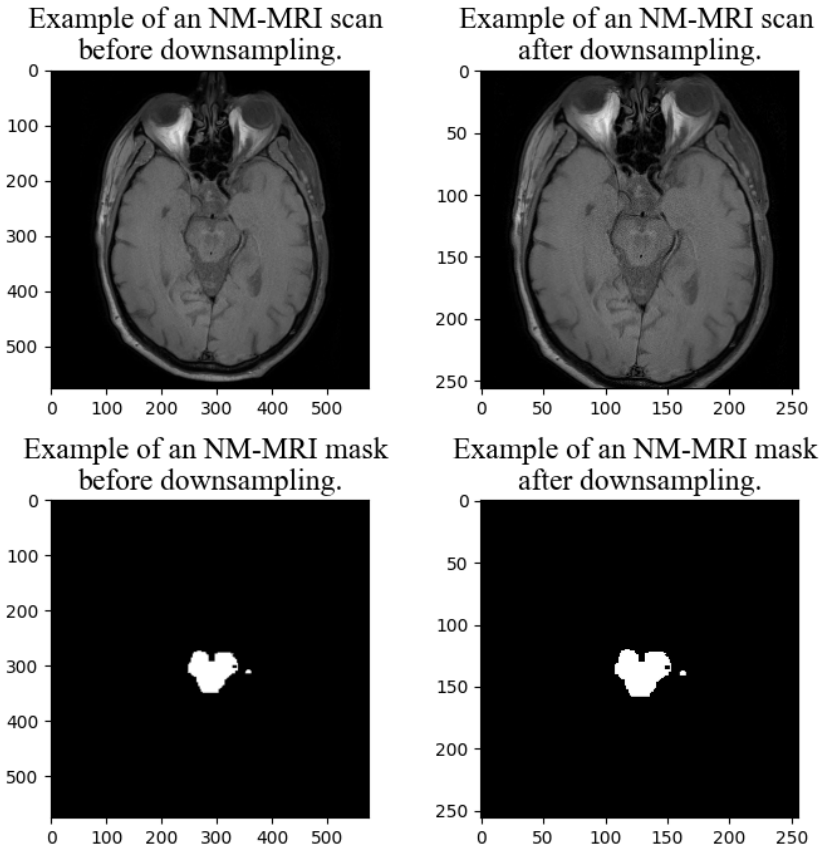


Figure 3.11. Example of an NM scan before the application of the downsampling operation and afterwards.

Table 3.5. Final division of the NM dataset considering training and testing sets after the data augmentation process.

		Original	Augmented	Total
NM subset	Training	468	4212	4680
	Testing	156	-	156
	Total	624	4212	4836

4 Results and Discussion

In this section, we present the segmentation results of both the T1W and NM images. The first case was a more systematic and profound study, which required an extensive grid search and defined the pre-training phase of the pipeline, with a larger dataset. The second case was an initial evaluation of the fine-tuned axial model, adjusted with NM images and previously trained with the large region subset of axial T1W scans.

4.1 T1W images

After the grid search optimization process, we selected the most efficient combinations of hyperparameters for the 4 models, considering both large and small region subsets, as shown in table 4.1. By most efficient, we mean the combination of hyperparameters that resulted in the best metrics with the smallest number of filters in the first layer. In Figure 4.1 it is possible to visualize some examples of each models' predictions.

Table 4.1. Most efficient combinations of hyperparameters for each model in both data subsets, considering the ones that resulted in the best metrics with the smallest number of filters in the first layer.

Data subset	Model	Optimizer	Learning rate	Number of filters in the first layer
Large region	Axial	AdamW	5e-4	32
	Sagittal	AdamW	5e-4	16
	Coronal	AdamW	5e-4	32
	Generic	AdamW	5e-4	16
Small region	Axial	Adam	5e-4	32
	Sagittal	Adam	1e-4	8
	Coronal	AdamW	5e-4	32
	Generic	AdamW	1e-4	16

We can highlight one interesting insight related to the number of filters in the first layers for axial and coronal based models in the large region subset. Both of them had their best performances with the highest possible number of filters in the first layer. This fact suggests that a higher number, such as 64, used in the original U-Net architecture, could possibly lead to even better segmentations.

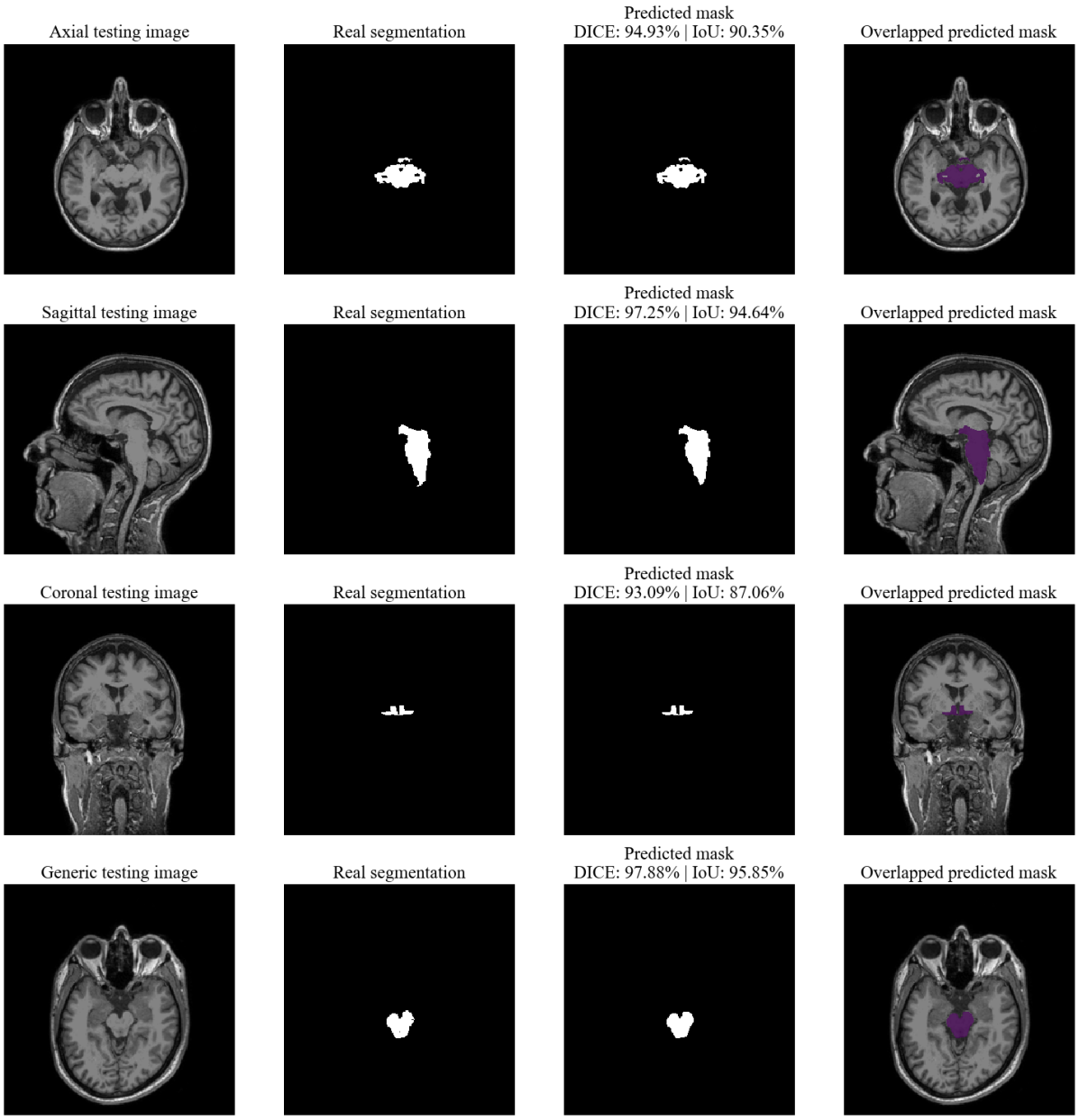


Figure 4.1. Examples of predictions made by each model where, in the first column, the original T1W image is exhibited, in the second, the ground truth mask, in the third, the segmentation predicted by the model and in the fourth, the overlapped predicted region on the original image. Now, the first row contains an example of prediction made by an axial model's on the large subset, the second, by a sagittal, the third, by a coronal and, lastly, the fourth, by a generic. Their respective DSC and IoU are also shown on the top of each predicted mask image.

It is also important to point out that cases with empty segmented set and also empty ground truth set were considered to have an IoU of 1. By definition, this metric would result in an indeterminate form with both numerator and denominator vanishing. Nevertheless, in the context of image segmentation, these are successful situations, where the model was able to predict correctly the absence of a ROI. Given that, table 4.2 exhibits the resulting metrics of the 4 models on the test set of both large and small datasets.

Table 4.2. Resultant metrics of all 4 models evaluated under both large region test set and small region test set, respectively, in terms of DSC, IoU, HD, SMHD and HD95. The slices used to test each model were the same as the ones they were trained with.

Data subset	Model	Mean DSC	Mean IoU	Mean HD	Mean SMHD	Mean HD95
Large region	Axial	93.88%	89.17%	1.89	0.09	0.70
	Sagittal	87.73%	81.87%	2.03	0.09	0.64
	Coronal	**95.34%**	**93.03%**	**0.68**	**0.04**	**0.27**
	Generic	92.42%	87.49%	1.95	0.10	0.74
Small region	Axial	88.99%	83.39%	1.88	0.14	1.07
	Sagittal	**96.68%**	**93.60%**	2.13	0.13	1.02
	Coronal	85.50%	80.52%	**1.44**	**0.07**	**0.50**
	Generic	89.91%	84.18%	1.86	0.14	1.06

4.2 Neuromelanin images

The combination of hyperparameters considered for the fine-tuning of the model were the ones that resulted in the best performance of the axial model trained with the large region subset. These are shown in Table 4.3. No grid search or any other type of hyperparameter tuning was used in this analysis.

Table 4.3. Hyperparameters used to fine-tune with NM scans the axial model trained with the large region subset. They are the same

Hyperparameter	Value
Optimizer	AdamW
Loss	Binary cross-entropy
Activation function	ReLU
Learning rate	5e-4
Number of filters in the first layer	32
Epochs	300
Patience	10
Batch size	1

Once the model was trained, it was applied to the test set and the average metrics were computed. In the same way as done with the T1W training, DSC was the main

metric used during training, while the others were considered to support the evaluation process. Thereby, the final metrics are shown in Table 4.4 and 4 examples of predictions are exhibited in Figure 4.2.

Table 4.4. Final metrics measuring the performance of the fine-tuned axial model with NM scans. The axial model's best performance hyperparameters trained with the large region subset were considered for this first analysis and the metrics used were the average of DSC, IoU, HD, SMHD, and HD95. The same division of subjects in the training and test sets were maintained, which means that the scans used to test the model were not previously presented to it, not even the T1W.

Dataset	Model	Mean DSC	Mean IoU	Mean HD	Mean SMHD	Mean HD95
NM scans	Fine-tuned axial	84.75%	76.26%	3.35	0.26	0.0

Considering all pixels from all images in the test set, we built two confusion matrices, presented in Figures 4.3 and 4.4, where the first one contains absolute numbers of pixels, and the second normalized values in terms of the predicted labels. The labels 0 and 1 indicate pixels that are not part or that are not classified as part of the brainstem by the model and the ones that are, respectively. Figure 4.3 evidences the imbalanced distribution of classes considering the background pixels, which are correspondent to more than 98%. On the other hand, Figure 4.4 highlights the percentages of correct and incorrect predictions in each group of predicted label.

This preliminary study showed that our approach has great chances of succeeding. Despite the average metrics being below those of the axial model trained with T1W, we noticed a pattern in the model's prediction. Most samples with the best segmentations were cases where the brainstem appeared completely and with its characteristic shape similar to Mickey Mouse. The first and third rows of Figure 4.2 illustrate these scenarios. Conversely, the scans with the lowest DSC and IoU metrics were the ones where the brainstem was not present, however, the ground truth masks indicated the opposite, as shown in the second and fourth rows for Figure 4.2.

This analysis evidences both an achievement and a limitation of this work. The achievement refers to the model's ability to recognize scans where the brainstem in fact appears. However, it also highlights a limitation of this study related to the semi-automatic generation of ground truth masks. As mentioned in Subsection 3.1.1, the region between the mesencephalon and the diencephalon is a miscellaneous of structures called Ventral "DC". Therefore, some scans present ground truth masks in regions that do not have brainstem. In spite of this constraint, the model showed higher sensitivity and precision in scans where the brainstem was present than in those that it was not, indicating that the model potentially learned how to recognize the brainstem.

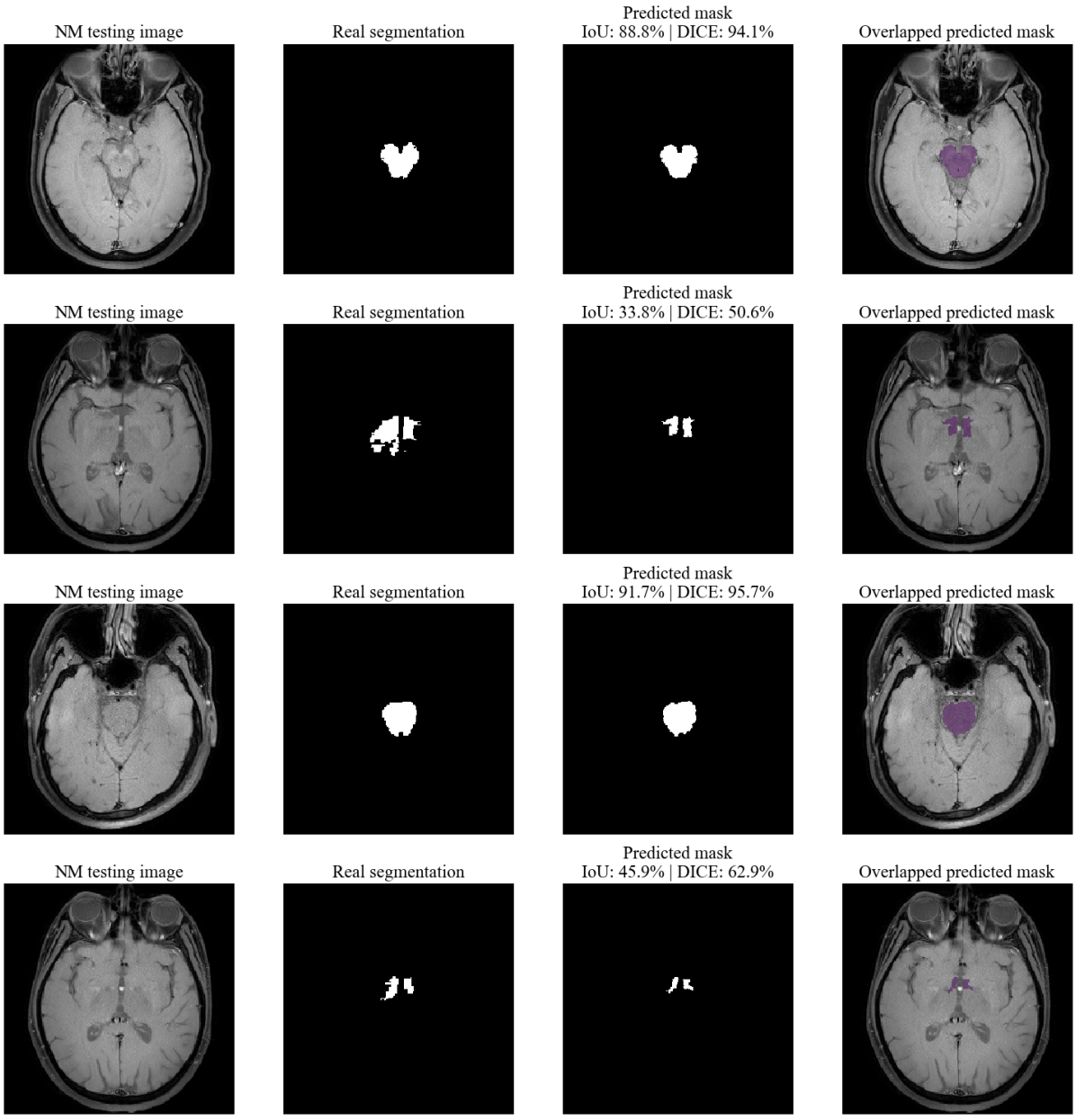


Figure 4.2. 4 different example cases predicted by the fine-tuned model. To the left, the input image is shown. Next to it, the ground truth masks. In the third column, the prediction. And, finally, in the last column, the input scan overlapped with the predicted mask.

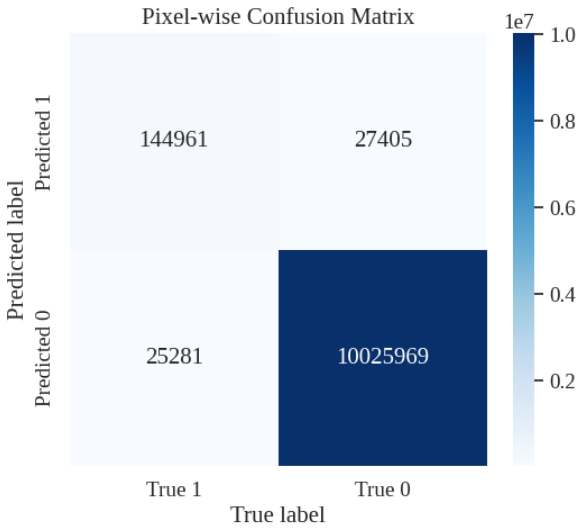


Figure 4.3. Confusion matrix with absolute numbers of pixels in all of test images, where 0 refers to the pixels that are not part or not predicted as part of the brainstem and 1 to the ones that are.

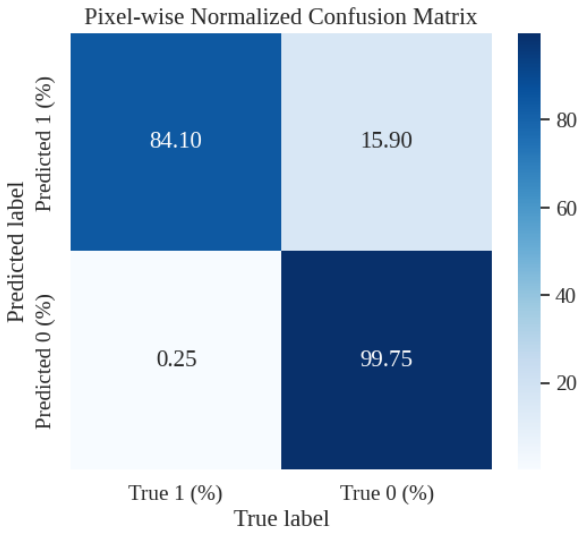


Figure 4.4. Confusion matrix normalized by each predicted label. In other words, the distribution of the correct and incorrect predictions among all pixels classified by the model as 0 (not part of the brainstem) or 1 (part of the brainstem).

4.3 Statistical Analysis

In order to evaluate the obtained results statistically, we first hypothesized normal distribution of the dataset. Through Lilliefors test, we rejected the null hypothesis of normality and, therefore, used the paired Wilcoxon test to compare the median of both DSC and IoU for each pair of models and then, these metrics to a fixed value.

The null hypothesis for the Wilcoxon test considered that there was no statistical difference between the metrics' medians from the generic model and the other model analyzed. In light of that, the results of the statistical analysis are shown in table 4.5.

Table 4.5. Results of the Wilcoxon statistical analysis of the models' DSC, IoU and HD medians. The highlighted terms indicate significant statistical difference, with a p-value of less or equal to 1%.

Data subset	Model compared	Upper median DSC (p-value)	Upper median IoU (p-value)	Lower median HD (p-value)
Large region	Axial Sagittal Coronal	**Axial** **Generic** **Coronal**	**Axial** **Generic** **Coronal**	Axial ($p < 0.03$) - ($p = 0.26$) **Coronal**
Small region	Axial Sagittal Coronal	- ($p = 0.55$) **Sagittal** - ($p = 0.89$)	- ($p = 0.55$) **Sagittal** - ($p = 0.88$)	- ($p = 0.35$) Generic ($p = 0.02$) **Coronal**

Thereafter, we compared all models trained with the large subset to fixed discrete values. With that, we can affirm that the DSC median of the axial model was significantly greater than 90% ($p < 0.01$), sagittal than 87% ($p < 0.01$), coronal than 95% ($p < 0.01$), and generic than 92% ($p < 0.01$).

In the large region subset, coronal model stood out, presenting significant statistical difference ($p < 0.01$) in all 3 metrics analyzed when compared to the generic model's performance. On the other hand, in the small region subset, sagittal model showed statistical significance ($p < 0.01$) in 2 out of 3 metrics. Still, the generic model's statistically significant results compared to the sagittal model ($p < 0.01$) are intriguing, considering the higher level of complexity due to the greater variability of shapes and locations of the brainstem, particularly in the large subset.

In this context, taking into account the comparable results of the axial model in the large region subset and the potential of its improvement with a greater number of filters in the first layer, our plan to apply these models to NM scans and expect commensurable outcomes stands firm. This relies on the fact that the anatomical plane that is closest to the NM oblique acquisition is the axial.

4.4 Limitations of the Work

4.4.1 General Limitations

The semi-automated process to create the ground truth masks brings a couple advantages in terms of time saving, scalability and reproducibility. However, considering that the gold standard is manual annotation, the use of Freesurfer's technology, even allied to the specialist's analysis, provided miscellaneous masks from the region between the midbrain and the diencephalon, as mentioned previously. Evidences of this limitation were raised in the case study of the fine-tuning process of the axial model applied to NM images.

Additionally, the restricted computational resources also contributed to the limited number of images in our dataset due to the long processing time. This limitation leads to an increased vulnerability of the learning process to the patient's positioning in the MRI machine. Since we chose a specific range of slices to compose the subsets, the narrower the range, the higher the chances to obtain slices without the ROI, which may impair the model's learning process.

4.4.2 Possible Biases

It is important to consider the possible biases is this study. Firstly, the intrinsic characteristic of a much higher number of true negative pixels in comparison to the true positive in the scans already provides imbalanced classes for the model to learn. However, this case was already addressed since we used specific metrics that measure precision and recall indirectly, such as DSC. Other two biases are related to the dataset composition. One is the fact that there are more male participants than female, which may generate a gender bias due to the fact that, in general, their anatomical dimensions of the brain and its structures are different. The other one refers to the higher number of PD patients compared to the number of CG subjects. This may lead to a bias because of the differences in terms of tissue composition and, therefore, contrast in the scans.

To address this issue properly, a possible approach would be to separate in subsets the participants we suspect that may be biased, for example, a group of women and a group of men. After that, we would apply the model in each group separately and compare the results, especially in terms of a confusion matrix in a pixel level. This way, it would be viable to compare the error rates of false positives and false negatives across groups, which is a widely used method to detect systematic bias in ML models [10].

This analysis would provide more robustness to the study since it verifies the model's generalization ability and grants it with transparency. Moreover, regarding the medical context, bias evaluation ensures that interventions based on this work will not disadvantage specific populations, since there may be other types of bias that we are not aware of. Also, an important interpretation of these results are related to the false positive and false negative rates, since the first one indicates unnecessary treatments and the second, missed diagnoses.

5 Conclusion

In this work, we present a comparative analysis between U-Net adjusted architectures trained with 2 different sized data subsets of distinct anatomical planes T1W scans, performing a segmentation task on the brainstem structure. In total, 4 final models were evaluated under DSC, IoU, HD, SMHD, and HD95 metrics: the axial model, the sagittal, the coronal, and the called generic, for being trained with all types of slices.

The definition of the best hyperparameters already raised insights indicating a probable potential to achieve better results with a higher number of filters in the first layer of the axial and coronal models trained with the large subset.

We emphasized the performance of the axial (DSC: 93.88%, IoU: 89.17%, HD: 1.89, SMHD: 0.09, HD95: 0.70) and coronal (DSC: 95.34%, IoU: 93.03%, HD: 0.68, SMHD: 0.04, HD95: 0.74) models in the large region subset. In the small region subset, the results were not as homogeneous as they were in the large, but still, the sagittal model can be highlighted in terms of DSC and IoU (DSC: 96.68%, IoU: 93.60%, HD: 2.13, SMHD: 0.13, HD95: 1.02), while, considering HD and its variations, the coronal model outstanded (DSC: 85.50%, IoU: 80.52%, HD: 1.44, SMHD: 0.77, HD95: 0.50). Also, in the large region subset, regarding the DSC medians, the axial showed to have a median significantly greater than 90% ($p < 0.01$), the sagittal than 87% ($p < 0.01$), the coronal than 95% ($p < 0.01$), and the generic than 92% ($p < 0.01$).

Additionally, we made a preliminary analysis of the axial model trained with the large subset and fine-tuned with NM scans to segment the brainstem. This case study raised relevant insights and expectations related to the transfer learning and the model's ability to detect the brainstem in NM images.

It is important to point out that we consider this study a first step towards a Computer-Aided Diagnosis (CAD) tool. Among all benefits an instrument like this could bring, we can mention the assistance to specialists to deliver earlier and more accurate diagnosis and the possibility to provide insights when analyzing the most relevant features the model relied on to determine its outputs with the use of explainable AI, for instance.

5.1 Future Work

Taking our work's limitations into consideration, we aim to generate gold standard ground truth masks, so that we can also include Freesurfer's segmentations and compare it to the other methods. In addition, we aspire to analyze the impact of the threshold value on the final results, as presented by [30]. Similarly, another approach that caught our attention was the one suggested by [59], who proposed a unilateral evaluation of the hemispheres of the brain, classifying them separately.

In addition, the results of our preliminary analysis of the fine-tuned model showed potential to execute the task of segmenting the brainstem structure out of scans acquired in the NM space. Therefore, as future work, we aim to analyze more deeply the hyperparameters choice of the pre-trained model, verifying if it reached its optimal performance peak. In the same way, we also would like to investigate the possible biases in the study to obtain an even more reliable work.

Bearing in mind that our final goal is to be able to quantify the NM signal intensity using 2 segmentation models, one to segment the brainstem, which is what this study covers mostly, followed by another one to segment the SN. We made a case study, applying the axial model to NM sensitive MRI scans to verify if its performance is maintained in this different space. Regarding the fewer number of this type of acquisitions, a model pre-trained in T1W showed potential to catalyze the learning process of the midbrain structure and provide promising results. Finally, if the quantification of the NM is successful, we can input it into a classifier which outputs may be binary: healthy or not, or result in a multiclass classification, determining the predicted disease stage.

References

[1] Mikel Ariz, Martín Martínez, Ignacio Alvarez, Maria A. Fernández-Seara, Gabriel Castellanos, Catalonian Neuroimaging Parkinson's Disease Consortium, Pau Pastor, Maria A. Pastor, and Carlos de Solórzano. Automatic segmentation and quantification of nigrosome-1 neuromelanin and iron in MRI: a candidate biomarker for Parkinson's disease. *Journal of Magnetic Resonance Imaging*, 60(2):534–547, 2024. Publisher: Wiley Online Library.

[2] Gopalan Balachandran. *MRI Brain: Atlas and Text*. Jaypee Brothers Medical Publishers (P) Ltd., 2016.

[3] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, and et al. The Liver Tumor Segmentation Benchmark (LiTS). *Medical Image Analysis*, 84:102680, February 2023.

[4] Benjamin Billot, Martina Bocchetta, Emily Todd, Adrian V. Dalca, Jonathan D. Rohrer, and Juan Eugenio Iglesias. Automated segmentation of the hypothalamus and associated subunits in brain MRI. *NeuroImage*, 223:117287, December 2020.

[5] Peter S Bloomfield, Sabrina Brigadoi, Gaia Rizzo, and Mattia Veronese. *Basic Neuroimaging: A guide to the methods and their applications*. CreateSpace, 2 edition, 2021.

[6] Heiko Braak, Kelly Del Tredici, Udo Rüb, Rob A.I De Vos, Ernst N.H Jansen Steur, and Eva Braak. Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of Aging*, 24(2):197–211, March 2003.

[7] Pedro Renato de Paula Brandão. *Comprometimento cognitivo na doença de Parkinson: correlatos clínicos, neuropsicológicos e de neuroimagem*. PhD Thesis, Universidade de Brasília, Brasília, DF, Brazil, 2022.

[8] Pedro Renato P. Brandão, Renato Puppi Munhoz, Talyta Cortez Grippe, Francisco Eduardo Costa Cardoso, Brenda Macedo De Almeida E Castro, Ricardo Titze-de

Almeida, Carlos Tomaz, and Maria Clotilde Henriques Tavares. Cognitive impairment in Parkinson's disease: A clinical and pathophysiological overview. *Journal of the Neurological Sciences*, 419:117177, December 2020.

[9] Gabriel Castellanos, María A. Fernández-Seara, Oswaldo Lorenzo-Betancor, Sara Ortega-Cubero, Marc Puigvert, Javier Uranga, Marta Vidorreta, Jaione Irigoyen, Elena Lorenzo, Arrate Muñoz-Barrutia, Carlos Ortiz-de-Solorzano, Pau Pastor, and María A. Pastor. Automated Neuromelanin Imaging as a Diagnostic Biomarker for Parkinson's Disease. *Movement Disorders*, 30(7):945–952, June 2015.

[10] Simon Caton and Christian Haas. Fairness in Machine Learning: A Survey. *ACM Computing Surveys*, 56(7):1–38, July 2024.

[11] K. Ray Chaudhuri, Jean-Philippe Azulay, Per Odin, Susanna Lindvall, Josefa Domingos, Ali Alobaidi, Prasanna L. Kandukuri, Vivek S. Chaudhari, Juan Carlos Parra, Toru Yamazaki, Julia Oddsdottir, Jack Wright, and Pablo Martinez-Martin. Economic Burden of Parkinson's Disease: A Multinational, Real-World, Cost-of-Illness Study. *Drugs - Real World Outcomes*, 11(1):1–11, March 2024.

[12] Max Dünnwald, Philipp Ernst, Emrah Düzel, Klaus Tönnies, Matthew J. Betts, and Steffen Oeltze-Jafra. Fully automated deep learning-based localization and segmentation of the locus coeruleus in aging and Parkinson's disease using neuromelanin-sensitive MRI. *International Journal of Computer Assisted Radiology and Surgery*, 16(12):2129–2135, December 2021.

[13] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, June 2006.

[14] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, August 2012.

[15] Bruce Fischl, David H. Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, Shuna Klaveness, Albert Montillo, Nikos Makris, Bruce Rosen, and Anders M. Dale. Whole Brain Segmentation. *Neuron*, 33(3):341–355, January 2002.

[16] Rahul Gaurav, Romain Valabrègue, Lydia Yahia-Chérif, Graziella Mangone, Sridar Narayanan, Isabelle Arnulf, Marie Vidailhet, Jean-Christophe Corvol, and Stéphane Lehéricy. NigraNet: An automatic framework to assess nigral neuromelanin content in early Parkinson's disease using convolutional neural network. *NeuroImage: Clinical*, 36:103250, 2022.

[17] Rafael C. Gonzalez and Richard E. Woods. *Digital image processing*. Prentice-Hall, Upper Saddle River, NJ, 2. ed., internat. ed edition, 2002.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

[19] Takashi Hashido and Shigeyoshi Saito. Quantitative T1, T2, and T2* Mapping and Semi-Quantitative Neuromelanin-Sensitive Magnetic Resonance Imaging of the Human Midbrain. *PLOS ONE*, 11(10):e0165160, October 2016.

[20] Naying He, Yongsheng Chen, Peter A. LeWitt, Fuhua Yan, and E. Mark Haacke. Application of Neuromelanin MR Imaging in Parkinson Disease. *Journal of Magnetic Resonance Imaging*, 57(2):337–352, February 2023.

[21] Margareth. M. Hoehn and Melvin. D. Yahr. Parkinsonism: onset, progression and mortality. *Neurology*, 17(5):427–442, May 1967.

[22] Institute for Health Metrics and Evaluation (University of Washington, USA). Brain Health Atlas. Available at https://brainhealthatlas.org/data/brain-health/line. The plot we used is obtained selecting the following conditions: 3.1 Headache disorders, 3.2 Other neurological disorders, 3.3 Motor neuron disease, 3.4 Multiple sclerosis, 3.5 Epilepsy, 3.6 Parkinson's disease, 3.7 Alzheimer's disease and other dementias. Also, the observed measured corresponds to the number of Deaths, and the corresponding selected location is the Global option. Last access: 01/18/2026.

[23] Mina Jafari, Ruizhe Li, Yue Xing, Dorothee Auer, Susan Francis, Jonathan Garibaldi, and Xin Chen. FU-Net: Multi-class Image Segmentation Using Feedback Weighted U-Net. In Yao Zhao, Nick Barnes, Baoquan Chen, Rüdiger Westermann, Xiangwei Kong, and Chunyu Lin, editors, *Image and Graphics*, volume 11902, pages 529–537. Springer International Publishing, Cham, 2019. Series Title: Lecture Notes in Computer Science.

[24] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani, editors. *An introduction to statistical learning: with applications in R*. Number 103 in Springer texts in statistics. Springer, New York, 2013.

[25] Eric R. Kandel. *The disordered mind: what unusual brains tell us about ourselves*. Farrar, Straus and Giroux, New York, first edition edition, 2018.

[26] Mahsa Karami, Pantea Majma Sanaye, Atousa Ghorbani, Roshanak Amirian, Pouya Goleij, Mehregan Babamohamadi, and Zhila Izadi. Recent advances in targeting LRRK2 for Parkinson's disease treatment. *Journal of Translational Medicine*, 23(1):754, July 2025.

[27] Kenichi Kashihara, Takayoshi Shinya, and Fumiyo Higaki. Neuromelanin magnetic resonance imaging of nigral volume loss in patients with Parkinson's disease. *Journal of Clinical Neuroscience*, 18(8):1093–1096, August 2011.

[28] Salman Khan, Hossein Rahmani, Syed Afaq Ali Shah, and Mohammed Bennamoun. *A Guide to Convolutional Neural Networks for Computer Vision*. Synthesis Lectures on Computer Vision. Springer International Publishing, Cham, 2018.

[29] Valliappa Lakshmanan, Martin Goerner, and Ryan Gillard. *Practical machine learning for computer vision: end-to-end machine learning for images*. O'Reilly, Bejing Boston Farnham Sebastopol Tokyo, first edition, second release edition, 2021.

[30] Alice Le Berre, Koji Kamagata, Yujiro Otsuka, Christina Andica, Taku Hatano, Laetitia Saccenti, Takashi Ogawa, Haruka Takeshige-Amano, Akihiko Wada, Michimasa Suzuki, Akifumi Hagiwara, Ryusuke Irie, Masaaki Hori, Genko Oyama, Yashushi Shimo, Atsushi Umemura, Nobutaka Hattori, and Shigeki Aoki. Convolutional neural network-based segmentation can help in assessing the substantia nigra in neuromelanin MRI. *Neuroradiology*, 61(12):1387–1395, December 2019.

[31] Mengyu Li, Magnús Magnússon, Ingibjörg Kristjánsdóttir, Sigrún Helga Lund, Thilo Van Eimeren, and Lotta M. Ellingsen. Region-based U-nets for fast, accurate, and scalable deep brain segmentation: Application to Parkinson Plus Syndromes. *NeuroImage: Clinical*, 47:103807, 2025.

[32] Nikos Makris, Marlene Oscar-Berman, Sharon Kim Jaffin, Steven M. Hodge, David N. Kennedy, Verne S. Caviness, Ksenija Marinkovic, Hans C. Breiter, Gregory P. Gasic, and Gordon J. Harris. Decreased Volume of the Brain Reward System in Alcoholism. *Biological Psychiatry*, 64(3):192–202, August 2008.

[33] Yasunari Matsuzaka and Ryu Yashiro. The Diagnostic Classification of the Pathological Image Using Computer Vision. *Algorithms*, 18(2):96, February 2025.

[34] Elisa Menozzi and Anthony HV Schapira. Prospects for disease slowing in parkinson disease. *Annual Review of Pharmacology and Toxicology*, 65, 2025.

[35] Nicola Mercuri and Giorgio Bernardi. The 'magic' of L-dopa: why is it the gold standard Parkinson's disease therapy? *Trends in Pharmacological Sciences*, 26(7):341–344, July 2005.

[36] Tom M. Mitchell. *Machine learning*. McGraw-Hill series in Computer Science. McGraw-Hill, New York, nachdr. edition, 2013.

[37] Dominik Müller, Iñaki Soto-Rey, and Frank Kramer. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1):210, December 2022.

[38] Salvatore Nigro, Marco Filardi, Benedetta Tafuri, Martina Nicolardi, Roberto De Blasi, Alessia Giugno, Valentina Gnoni, Giammarco Milella, Daniele Urso, Stefano Zoccolella, Giancarlo Logroscino, for the Frontotemporal Lobar Degeneration Neuroimaging Initiative, for the 4-Repeat Tau Neuroimaging Initiative, and for the Alzheimer's Disease Neuroimaging Initiative. Deep Learning–based Approach for Brainstem and Ventricular MR Planimetry: Application in Patients with Progressive Supranuclear Palsy. *Radiology: Artificial Intelligence*, 6(3):e230151, May 2024.

[39] José A. Obeso, Maria Stamelou, Christopher G. Goetz, Werner Poewe, Anthony E. Lang, Daniel Weintraub, David Burn, Glenda M. Halliday, Erwan Bezard, Serge Przedborski, Stéphane Lehericy, David J. Brooks, John C. Rothwell, Mark Hallett, Mahlon R. DeLong, Connie Marras, Caroline M. Tanner, Gary W. Ross, J. William Langston, Christine Klein, Vincenzo Bonifati, Joseph Jankovic, Andres M. Lozano, Günther Deuschl, Hagai Bergman, Eduardo Tolosa, Martha Rodriguez-Violante, Stanley Fahn, Ronald B. Postuma, Daniela Berg, Kenneth Marek, David G. Standaert, D. James Surmeier, C. Warren Olanow, Jeffrey H. Kordower, Paolo Calabresi, Anthony H. V. Schapira, and Anthony J. Stoessl. Past, present, and future of Parkinson's disease: A special essay on the 200th Anniversary of the Shaking Palsy. *Movement Disorders*, 32(9):1264–1310, September 2017.

[40] James Parkinson. An Essay on the Shaking Palsy. *The Journal of Neuropsychiatry and Clinical Neurosciences*, 14(2):223–236, May 2002.

[41] Aslam Pathan and Abdulrahman Alshahrani. Gold Standard of Symptomatic treatment in Parkinson disease: Carbidopa / Levodopa. *NeuroPharmac Journal*, pages 63–68, December 2018.

[42] Shaoyi Peng, Peng Liu, Xiaowen Wang, and Kaiyuan Li. Global, regional and national burden of Parkinson's disease in people over 55 years of age: a systematic analysis of the global burden of disease study, 1991–2021. *BMC Neurology*, 25(1):178, April 2025.

[43] Jannik Prasuhn, Michelle Prasuhn, Anja Fellbrich, Robert Strautz, Felicitas Lemmer, Shalida Dreischmeier, Meike Kasten, Thomas F. Münte, Henrike Hanssen, Marcus Heldmann, and Norbert Brüggemann. Association of Locus Coeruleus and Substantia Nigra Pathology With Cognitive and Motor Functions in Patients With Parkinson Disease. *Neurology*, 97(10), September 2021.

[44] Thomas C. Pritchard and Kevin D. Alloway. *Medical neuroscience*. Fence Creek Pub. ; Distributors, U.S. and Canada, Blackwell Science, Madison, Conn., Malden, MA, 1st ed edition, 1999. OCLC: 41086829.

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241, 2015.

[46] Laura Sander, Simon Pezold, Simon Andermatt, Michael Amann, Dominik Meier, Maria J. Wendebourg, Tim Sinnecker, Ernst-Wilhelm Radue, Yvonne Naegelin, Cristina Granziera, Ludwig Kappos, Jens Wuerfel, Philippe Cattin, Regina Schlaeger, and Alzheimer's Disease Neuroimaging Initiative. Accurate, rapid and reliable, fully automated MRI brainstem segmentation for application in multiple sclerosis and neurodegenerative diseases. *Human Brain Mapping*, 40(14):4091–4104, October 2019.

[47] Makoto Sasaki, Eri Shibata, Koujiro Tohyama, Junko Takahashi, Kotaro Otsuka, Kuniaki Tsuchiya, Satoshi Takahashi, Shigeru Ehara, Yasuo Terayama, and Akio Sakai. Neuromelanin magnetic resonance imaging of locus ceruleus and substantia nigra in Parkinson's disease. *NeuroReport*, 17(11):1215–1218, July 2006.

[48] Stefan T. Schwarz, Timothy Rittman, Vamsi Gontu, Paul S. Morgan, Nin Bajaj, and Dorothee P. Auer. T1-Weighted MRI shows stage-dependent substantia nigra signal loss in Parkinson's disease: Substantia Nigra T1 Signal Loss In PD. *Movement Disorders*, 26(9):1633–1638, August 2011.

[49] Larry J. Seidman, Stephen V. Faraone, Jill M. Goldstein, Julie M. Goodman, William S. Kremen, Genichi Matsuda, Elizabeth A. Hoge, David Kennedy, Nikos Makris, Verne S. Caviness, and Ming T. Tsuang. Reduced subcortical brain volumes in nonpsychotic siblings of schizophrenic patients: A pilot magnetic resonance imaging study. *American Journal of Medical Genetics*, 74(5):507–514, September 1997.

[50] Sanjif Shanmugavelu, Mathieu Taillefumier, Christopher Culver, Oscar Hernandez, Mark Coletti, and Ada Sedova. Impacts of floating-point non-associativity on reproducibility for HPC and deep learning applications. In *SC24-W: Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, pages 170–179, Atlanta, GA, USA, November 2024. IEEE.

[51] Khushi Sharma, Manjula Shanbhog, and Kuljeet Singh. Machine learning in neuroimaging and computational pathophysiology of Parkinson's disease: A comprehensive review and meta-analysis. *Asian Journal of Psychiatry*, 109:104537, July 2025.

[52] Sahar Shokrpour, AmirMehdi MoghadamFarid, Sepideh Bazzaz Abkenar, Mostafa Haghi Kashani, Mohammad Akbari, and Mostafa Sarvizadeh. Machine learning for

Parkinson's disease: a comprehensive review of datasets, algorithms, and challenges. *npj Parkinson's Disease*, 11(1):187, July 2025.

[53] Jaimie D Steinmetz, Katrin Maria Seeher, Nicoline Schiess, Emma Nichols, Bochen Cao, Chiara Servili, and et al. Global, regional, and national burden of disorders affecting the nervous system, 1990–2021: a systematic analysis for the Global Burden of Disease Study 2021. *The Lancet Neurology*, 23(4):344–381, April 2024.

[54] David Sulzer, Clifford Cassidy, Guillermo Horga, Un Jung Kang, Stanley Fahn, Luigi Casella, Gianni Pezzoli, Jason Langley, Xiaoping P. Hu, Fabio A. Zucca, Ioannis U. Isaias, and Luigi Zecca. Neuromelanin detection by magnetic resonance imaging (MRI) and its promise as a biomarker for Parkinson's disease. *npj Parkinson's Disease*, 4(1):11, April 2018.

[55] Iván Sánchez Fernández and Jurriaan M. Peters. Machine learning and deep learning in medicine and neuroimaging. *Annals of the Child Neurology Society*, 1(2):102–122, June 2023.

[56] Avner Thaler. Structural and Functional MRI in Familial Parkinson's Disease. In *International Review of Neurobiology*, volume 142, pages 261–287. Elsevier, 2018.

[57] Ricardo Titze-de Almeida, Simoneide Souza Titze-de Almeida, Gabriel Ginani Ferreira, Andrezza Paula Brito Silva, Pedro Renato De Paula Brandão, Wolfgang H. Oertel, Carlos H. Schenck, and Raimundo Nonato Delgado Rodrigues. microRNA signatures in prodromal REM sleep behavior disorder and early Parkinson's disease as noninvasive biomarkers. *Sleep Medicine*, 78:160–168, February 2021.

[58] Yida Wang, Naying He, Chunyan Zhang, Youmin Zhang, Chenglong Wang, Pei Huang, and et al. An automatic interpretable deep learning pipeline for accurate Parkinson's disease diagnosis using quantitative susceptibility mapping and T1-weighted images. *Human Brain Mapping*, 44(12):4426–4438, August 2023.

[59] Thomas Welton, Septian Hartono, Weiling Lee, Peik Yen Teh, Wenlu Hou, Robert Chun Chen, Celeste Chen, Ee Wei Lim, Kumar M. Prakash, Louis C. S. Tan, Eng King Tan, and Ling Ling Chan. Classification of Parkinson's disease by deep learning on midbrain MRI. *Frontiers in Aging Neuroscience*, 16:1425095, August 2024.

[60] Graham Winston, Natasha Kharas, Per Svenningsson, Ashwani Jha, and Michael G Kaplitt. Gene therapy for Parkinson's disease: trials and technical advances. *The Lancet Neurology*, 24(6):548–556, June 2025.

[61] World Health Organization. *Parkinson Disease: a Public Health Approach. Technical Brief*. World Health Organization, Geneva, 1st ed edition, 2022.

[62] Bin Xiao and Eng-King Tan. Prasinezumab slows motor progression in Parkinsons disease: beyond the clinical data. *npj Parkinson's Disease*, 11(1):31, February 2025.

[63] Varduhi Yeghiazaryan and Irina Voiculescu. An overview of current evaluation methods used in medical image segmentation. Technical Report RR-15-08, Department of Computer Science, Oxford, UK, 2015.

[64] Chiahui Yen, Chia-Li Lin, and Ming-Chang Chiang. Exploring the Frontiers of Neuroimaging: A Review of Recent Advances in Understanding Brain Functioning and Disorders. *Life*, 13(7):1472, June 2023.

[65] Luigi Zecca, Ruggero Fariello, Peter Riederer, David Sulzer, Alberto Gatti, and Davide Tampellini. The absolute concentration of nigral neuromelanin, assayed by a new sensitive method, increases throughout the life and is dramatically decreased in Parkinson's disease. *FEBS Letters*, 510(3):216–220, January 2002.

[66] Xi Jia Zhou, Chris Doyle, Logan Cross, Michael C Frank, and Nick Haber. Simulating variation in infant-caregiver attachment using reinforcement learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, 2025.