



UNIVERSITY
OF UDINE

Department of
Mathematics, Computer Science and Physics

PH.D. THESIS IN
COMPUTER SCIENCE, MATHEMATICS AND PHYSICS

Computer-aided analysis of complex neurological data for age-based classification of upper limbs motor performance and radiomics-based survival prediction of brain tumors

CANDIDATE

Asma Shaheen

SUPERVISOR

Dr. Stefano Burigat

CO-SUPERVISOR

Dr. Hassan Mohy UD Din

Cycle XXXIII — A.Y. 2021-2022

INSTITUTE CONTACTS

Dipartimento di Scienze Matematiche, Informatiche e Fisiche

Università degli Studi di Udine

Via delle Scienze, 206

33100 Udine — Italia

+39 0432 558400

<https://www.dmif.uniud.it/>

AUTHOR'S CONTACTS

shaheen.asma@spes.uniud.it

© 2022 Asma Shaheen

This work is shared under the Creative Commons 4.0 License Attribution-NonCommercial-ShareAlike.

Dedicated to my beautiful family for keeping their faith in me...!!

Abstract

Nowadays, the availability of an ever-increasing amount of digital medical data collected through heterogeneous sources such as healthcare systems, sensors, and mobile consumer technologies makes it possible to perform computer-aided analyses aimed at improving the knowledge, diagnosis, and treatment of medical conditions.

In this thesis, we worked with two medical datasets that can be used to study two different types of neurological disorders, motor control disorders (e.g., Parkinson's disease) and brain tumors.

The first dataset is comprised of the results of digital motor tests of the upper limbs that have been taken by more than 10000 users of a free and publicly available mobile application called MotorBrain. Motor tests are used by neurologists to assess human motor performance and support the diagnosis of disorders affecting motor control.

Our first goal was to analyse the MotorBrain data with statistical methods to investigate the age-related behavior patterns of healthy subjects for the different motor tests included in the application. Results show that the collected data reveal the typical degradation of motor performance that is common with aging, thus providing support for the appropriateness of the considered approach to motor performance data collection and potentially helping neurologist to identify neurological disorders at an early stage by comparing new data with the available normative data. At the same time, the results highlight problems that emerge when data collection is performed in an unsupervised non-clinical setting.

Based on the results of the statistical analysis, we used machine learning to automatically classify users according to their motor performance. The idea is to use such classification to automatically flag cases whose motor performance differs significantly from the typical performance of their age group and thus require manual inspection from a neurologist. In particular, we used random forest and logistic regression classification techniques with Minimum Redundancy, Maximum Relevance (MRMR) and Recursive Feature Elimination with SVM (RFE-SVM) feature selection methods. For each motor test, we were able to achieve good average accuracy in discriminating motor performance of young and old adults, with the random forest method leading to better results. Similar results were obtained for multi-class discrimination based on 5 age groups.

The second dataset we worked with consists of a standard set of MRI images of brain tumors that is often used to develop and validate radiomics-based methods for overall survival (OS) classification of brain gliomas. We specifically focused on two important steps of the radiomics process, segmentation and feature selection.

We first used the MRI dataset to empirically evaluate the impact of six different segmentation algorithms – five Convolutional Neural Networks and the STAPLE-fusion method - and four multiregional radiomic models (Whole Tumor (WT), 3-subregions, 6-subregions, and 21-subregions) on OS classification. Results of the evaluation show that the 3-subregions radiomics model has high predictive power but poor robustness while the 6-subregions and 21-subregions radiomics models are more robust but have low predictive power. The poor robustness of the 3-subregions radiomics model was associated with highly variable and inferior segmentation of tumor core and active tumor subregions as quantified by the Hausdorff metric. Failure analysis revealed that the WT radiomics model, the 6-subregions radiomics model, and the 21-subregions radiomics model failed for some subjects, possibly because of inaccurate segmentation of the WT volume. Moreover, short-term survivors were largely misclassified by the ra-

diomic models and were associated to large segmentation errors. The STAPLE fusion method was able to circumvent these segmentation errors but was not found to be the ultimate solution in terms of its predictive power.

We also evaluated the robustness of radiomic features with respect to automatic segmentation variability. To this purpose, we took into consideration seven state-of-the-art CNNs methods for brain tumor segmentation. We used the intra-class correlation coefficient (ICC) and overall correlation coefficient (OCCC) to quantitatively measure the robustness of radiomic features across the seven (independent) segmentation methods. We employed two feature selection techniques to select discriminatory features: Minimum Redundancy, Maximum Relevance (MRMR) and Recursive Feature Elimination with SVM (RFE-SVM). We then evaluated the effect of using robust radiomic features for OS classification by incorporating stability into feature selection methods, considering both stable features (via ICC and OCCC) and discriminatory features (via MRMR and RFE-SVM). Results show improvement in OS classification when using both stable and discriminatory features compared to using discriminatory features alone.

Declaration

I declare that the dissertation has been composed by myself and that the work has not been submitted for any other degree or professional qualification. I confirm that the work submitted is my own, except where work that has formed part of jointly-authored publications has been included. My contribution and those of the other authors to this work have been explicitly referenced. I confirm that appropriate credit has been given within this thesis where reference has been made to the work of others.

Name: Asma Shaheen

Date: November 19, 2022

List of Publications

- Journal publications

- **Asma Shaheen**, Syed Talha Bukhari, Maria Nadeem, Stefano Burigat, Ulas Bagci, Hassan Mohy-ud-Din. "*Overall Survival Prediction of Glioma Patients with Multiregional Radiomics*". *Frontiers in Neuroscience* 16 (2022).

- Conference publications

- **Asma Shaheen**, Stefano Burigat, Ulas Bagci, and Hassan Mohy-ud-Din. "*Overall survival prediction in gliomas using region-specific radiomic features.*" In *Machine Learning in Clinical Neuroimaging and Radiogenomics in Neuro-oncology*, pp. 259-267. Springer, Cham, 2020.

- Papers under submission

- **Asma Shaheen**, Stefano Burigat, Luca Chittaro, Riccardo Budai. "*Age-based analysis of a very large dataset of motor performance data collected in an unsupervised non-clinical setting.*" *Computer Methods and Programs in Biomedicine*.
- **Asma Shaheen**, Syed Talha Bukhari, Maria Nadeem, Stefano Burigat, Ulas Bagci, Hassan Mohy-ud-Din. "*Stable radiomic features with automatic segmentations variability for overall survival classification in glioblastoma.*"

Acknowledgements

I would like to thank my Ph.D. supervisor Dr. Stefano Burigat from the bottom of my heart for giving me the opportunity to work under his supervision and for always being kind to me during this journey. I would also like to thank my co-supervisor Dr. Hassan Mohy Ud for sparking my interest in the field of radiomics. His support, motivation, and guidance at every stage enabled me to complete my work successfully. Many thanks to Dr. Ulas Bagci, his mentoring has helped me to expand my knowledge and develop the skills that will help me in my career as a researcher.

Furthermore, I would also like to thank my loving husband, without whose affection, care, and constant efforts it would not have been possible for me to sustain this phase of research for a long period, and also my son Musa, for making me feel relaxed with his smile and presence. Many thanks to my sister Saira for her constant support throughout this journey. Thanks to my friend Pankaj Mishra, for supporting me when needed and being part of this journey.

Finally, words cannot describe how grateful I am to my loving and affectionate father and my sister Sobia. I am indebted to them for their constant support and for lifting me with their love in difficult situations in my life.

Contents

| | |
|---|-------------|
| Dedication | i |
| Abstract | ii |
| Declaration | v |
| List of Publications | vi |
| Contents | viii |
| List of Figures | xiii |
| List of Tables | xvi |
| 1 Introduction | 1 |
| 2 Background: digital motor tests for upper limbs motor performance assessment | 7 |
| 2.1 Motor control, movement disorders, and aging | 8 |
| 2.2 Tests of motor control | 9 |
| 2.3 Mobile apps for motor control testing | 11 |
| 2.4 The MotorBrain app | 14 |

| | | |
|----------|--|-----------|
| 2.4.1 | Motor tests | 14 |
| 2.4.2 | Recorded data | 19 |
| 2.5 | Our research goals | 21 |
| 3 | The MotorBrain dataset: overview and data cleaning process | 23 |
| 3.1 | Overview of the MotorBrain dataset | 24 |
| 3.2 | The data cleaning process | 26 |
| 3.2.1 | Small or erroneous screen size | 27 |
| 3.2.2 | Missing values | 28 |
| 3.2.3 | Incomplete runs | 29 |
| 3.2.4 | Anomalous psychophysical state of the user | 29 |
| 3.2.5 | Incorrect/incomplete performance | 29 |
| 3.3 | Results | 32 |
| 4 | Statistical analysis of the MotorBrain dataset | 34 |
| 4.1 | Performance measures | 35 |
| 4.1.1 | Trail making measures | 35 |
| 4.1.2 | Finger tapping measures | 37 |
| 4.2 | Young adults vs. old adults data analysis | 38 |
| 4.2.1 | Results | 40 |
| 4.2.2 | Discussion | 48 |
| 4.3 | Fine-grained age-based data analysis | 50 |
| 4.3.1 | Results | 52 |
| 4.3.2 | Discussion | 64 |
| 5 | Machine learning models for age-based classification of motor performance | 66 |
| 5.1 | Feature extraction | 67 |

| | | |
|----------|--|-----------|
| 5.1.1 | Trail making features | 67 |
| 5.1.2 | Finger tapping features | 69 |
| 5.2 | Feature selection | 71 |
| 5.2.1 | MRMR | 71 |
| 5.2.2 | RFE-SVM | 72 |
| 5.3 | Synthetic data generation | 72 |
| 5.4 | Classification | 73 |
| 5.4.1 | Random Forest | 73 |
| 5.4.2 | Logistic Regression | 74 |
| 5.5 | Experimental setup | 74 |
| 5.6 | Results | 75 |
| 5.6.1 | Feature selection | 75 |
| 5.6.2 | Classification results | 76 |
| 5.7 | Discussion | 80 |
| 6 | Background: brain tumor survival prediction with radiomics analysis | 84 |
| 6.1 | Gliomas | 85 |
| 6.2 | Medical imaging for brain tumor diagnosis and evaluation | 86 |
| 6.3 | Overall survival prediction | 87 |
| 6.4 | The radiomics process | 88 |
| 6.4.1 | Preprocessing | 89 |
| 6.4.2 | Segmentation | 89 |
| 6.4.3 | Feature extraction | 90 |
| 6.4.4 | Feature selection | 91 |
| 6.4.5 | Machine learning model development | 93 |
| 6.4.6 | Evaluation | 93 |
| 6.5 | State of the art on brain tumor segmentation | 94 |

| | | |
|----------|--|------------|
| 6.5.1 | CNN-based segmentation of brain tumor volume | 96 |
| 6.5.2 | Atlas-based segmentation | 100 |
| 6.6 | State-of-the-art on robustness of radiomic features | 102 |
| 6.7 | State of the art on radiomics-based OS prediction | 103 |
| 6.8 | Our research goals | 107 |
| 7 | Evaluation of the impact of segmentation algorithms on OS prediction with multiregional radiomics | 109 |
| 7.1 | Experimental methodology | 110 |
| 7.1.1 | Data | 110 |
| 7.1.2 | Preprocessing | 111 |
| 7.1.3 | Brain tumor segmentation | 112 |
| 7.1.4 | Tumor subregion segmentation models | 114 |
| 7.1.5 | Radiomic feature extraction | 116 |
| 7.1.6 | OS prediction: model training and inference | 118 |
| 7.1.7 | Evaluation | 118 |
| 7.2 | Results | 120 |
| 7.2.1 | Clinical characteristics | 120 |
| 7.2.2 | Segmentation algorithm performance | 121 |
| 7.2.3 | Radiomics models evaluation on Testing Cohort A (31 subjects) | 123 |
| 7.2.4 | Failure analysis | 125 |
| 7.2.5 | Radiomics models evaluation on Testing Cohort B (29 subjects) | 127 |
| 7.3 | Discussion | 128 |
| 8 | Identification of robust features and evaluation of their impact on OS prediction | 133 |
| 8.1 | Experimental methodology | 134 |
| 8.1.1 | Data | 134 |

| | | |
|-------|--|-----|
| 8.1.2 | Preprocessing | 134 |
| 8.1.3 | Brain tumor segmentation | 134 |
| 8.1.4 | Radiomic feature extraction | 136 |
| 8.1.5 | Stability analysis | 138 |
| 8.1.6 | Feature selection | 139 |
| 8.1.7 | OS prediction: model training and inference | 139 |
| 8.1.8 | Feature set reduction using prior information | 141 |
| 8.1.9 | Evaluation | 142 |
| 8.2 | Results | 142 |
| 8.2.1 | Clinical characteristics | 142 |
| 8.2.2 | Segmentation algorithm performance | 142 |
| 8.2.3 | Stability analysis results | 143 |
| 8.2.4 | Overall survival classification | 145 |
| 8.2.5 | Effect of feature selection based on prior information | 148 |
| 8.3 | Discussion | 151 |

9 Thesis conclusion and

| | |
|--------------------|------------|
| future work | 154 |
|--------------------|------------|

List of Figures

| | | |
|-----|--|----|
| 2.1 | Distribution of individuals with Parkinson’s disease by country from 2005 to 2030 [47] | 9 |
| 2.2 | MotorBrain tests are grouped in three categories based on the motor control measure they focus on. The Accuracy category is enabled in this example. The number on the right shows the number of times a users completed all the tests within that category. [167]. | 15 |
| 2.3 | MotorBrain tests designed to measure accuracy: (A) Circle-A: follow the colored ring starting from the highlighted position. (B) Square: follow the square frame starting from the highlighted position. | 16 |
| 2.4 | MotorBrain tests designed to measure Speed: (A) Circle-S: follow the ring as many times as possible within a time limit. (B) Path: follow the path once as fast as possible. | 18 |
| 2.5 | MotorBrain tests designed to measure Reaction Time: (A) Tapping2: tap as many times as possible on the active target, which is visible in one out of two possible positions and changes position after each tap. (B) Tapping4: tap as many times as possible on the active target, which is visible in one out of four possible positions and changes position after each tap. | 19 |
| 3.1 | Age distribution of users | 25 |

| | | |
|-----|---|-----|
| 3.2 | Examples of complete/correct vs incomplete/incorrect performance in the six motor tests. Each image shows all 3 repetitions of a test run with a different color. | 31 |
| 3.3 | Flow diagram of the cleaning process showing the results of each activity in terms of the number n of records removed in that step. | 33 |
| 4.1 | Mean values of performance measures in Circle-A task by age group | 57 |
| 4.2 | Mean values of performance measures in the Square test by age group | 58 |
| 4.3 | Mean values of performance measures in the Circle-S test by age group | 59 |
| 4.4 | Mean values of performance measures in the Path test by age group | 61 |
| 4.5 | Mean values of performance measures in the Tapping-2 test by age group | 62 |
| 4.6 | Mean values of performance measures in the Tapping-4 test by age group | 63 |
| 6.1 | Distribution of brain tumors in 2011-2015 in the United States [122]. | 86 |
| 6.2 | Feature selection is a multi-stage process consisting of removing unstable features, removing features with zero or near zero variance (non-informative features), removing highly correlated features (redundant features), selecting the optimal set of feature by using wrapper methods like maximum relevance minimum redundancy (mRMR) or recursive feature elimination (RFE) [123]. | 91 |
| 6.3 | Example input dataset with four MRI modalities and corresponding ground truth segmentation map. The last frame on the right is the ground truth with corresponding manual segmentation annotation. Label legend: enhancing tumor (green), peritumoral edema (yellow) and necrotic and non-enhancing tumor (red) [104]. | 94 |
| 7.1 | Automatically segmented tumor subregions from the five CNNs-based segmentation networks and the STAPLE fusion method. Label legend: Peritumoral Edema (green), Enhancing Core (yellow), Non-enhancing Core (orange). | 122 |

| | | |
|-----|--|-----|
| 7.2 | Distribution of misclassified subjects in (A) <i>WT</i> radiomics model, <i>3-subregions</i> radiomics model, and <i>6-subregions</i> radiomics model (B) <i>WT</i> radiomics model, <i>3-subregions</i> radiomics model, and <i>21-subregions</i> radiomics model (C) <i>WT</i> radiomics model, <i>6-subregions</i> radiomics model, and <i>21-subregions</i> radiomics model, on testing cohort A (31 subjects) | 126 |
| 8.1 | Distribution of stable features across different feature categories. | 144 |

List of Tables

| | | |
|-----|--|----|
| 3.1 | Demographics of MotorBrain users in the collected dataset. | 24 |
| 4.1 | Number of users by group for the different motor tests. | 39 |
| 4.2 | Descriptive statistics (mean and standard deviation (sd)) of the performance measures for the two age groups (young adults (18-30) and old adults (50-75)) | 41 |
| 4.3 | Main effects and interaction effects for age group and hand for the different performance measures. (Significant differences ($p < 0.05$) in bold) | 42 |
| 4.4 | Number of users by the group for the different motor tests. | 51 |
| 4.5 | Main effects and interaction effects for age group and hand for the different performance measures. (Significant differences ($p < 0.05$) in bold) | 53 |
| 4.6 | Pairwise comparisons between groups after significant interaction effects. Significant differences are highlighted in green. | 54 |
| 4.7 | Pairwise comparisons between dominant and non-dominant hand after significant interaction effects. Significant differences are highlighted in green. | 55 |
| 5.1 | Description of features extracted from shapes where {TMT}: trial making tests, {TT}: for tapping tests. | 70 |
| 5.2 | Top features selected by MRMR: 12 features for trail making tests and 6 for tapping tests ({TMT}: trial making tests, {TT}: for tapping tests). | 76 |

| | | |
|-----|--|-----|
| 5.3 | Classification accuracy (%) and AUCs of the RF and LR classifiers with the MRMR and RFE-SVM feature selection methods for motor performance classification (young vs. old adults) | 77 |
| 5.4 | AUCs of the RF and LR classifiers with the MRMR and RFE-SVM feature selection methods for Circle-A, Square, and Circle-S motor performance classification (5 Age groups): a and b refer to dominant and non-dominant hand results. | 79 |
| 5.5 | AUCs of the RF and LR classifiers with the MRMR and RFE-SVM feature selection methods for for Path, Tapping-2, and Tapping-4 motor performance classification (5 Age groups): a and b refer to dominant and non-dominant hand results. | 80 |
| 6.1 | Comparison of the state-of-the-art methods on the BraTS validation dataset (years 2018-2021). | 101 |
| 7.1 | Overview of the training and testing cohorts (A and B) used in the overall survival classification task | 112 |
| 7.2 | Configuration and hyperparameters of the five CNNs used for automatic segmentation of brain tumor volume (data provided by Syed. Talha Bukhari). . . | 113 |
| 7.3 | Summary of radiomic features extracted for four radiomic models, namely, <i>WT</i> radiomics model, <i>3-subregions</i> radiomics model, <i>6-subregions</i> radiomics model, and <i>21-subregions</i> radiomics model | 117 |
| 7.4 | Performance of the five CNNs-based segmentation networks and the STAPLE-fusion method on testing cohorts A and B (60 subjects). Bold font indicates best scores for overlapping subregions (WT, TC, and EC) | 123 |
| 7.5 | Performance of the 4 radiomic models on testing cohort A (31 subjects). Bold font indicates best performance achieved for each radiomic model. | 123 |

| | | |
|-----|---|-----|
| 7.6 | Classification accuracy of the four radiomics models on testing cohort B (29 subjects) | 127 |
| 7.7 | The 21 subregions defined by the Harvard-Oxford subcortical atlas | 129 |
| 8.1 | Configuration and hyperparameters of the seven CNNs used for automatic segmentation of brain tumor volume (data provided by Syed. Talha Bukhari). . . | 135 |
| 8.2 | The list of radiomic features extracted for each subject in the training and testing cohorts. | 137 |
| 8.3 | Performance of the considered segmentation schemes on testing cohort (125 subjects). | 143 |
| 8.4 | A summary of features selected to build the model for the overall survival classification task. | 146 |
| 8.5 | Model performance on the eight considered segmentation schemes on the testing cohort (31 subjects) | 148 |
| 8.6 | Model performance with the eight considered segmentation schemes on testing cohort (31 subjects) with priori selected features | 151 |

1

Introduction

The human nervous system is responsible for all actions of the human body, from walking to interacting with objects to sleeping. It is an advanced and complex network consisting of the brain, spinal cord, and nerves distributed throughout the body [162]. The billions of neurons in the brain communicate with the other parts of the body by transmitting and receiving electrical signals through the spinal cord and nerves. Many different neurological disorders can affect this system. Common disorders include epilepsy and seizures, stroke, acute spinal cord injury, brain tumors, and neurodegenerative diseases such as Parkinson's [106]. These disorders are a consequence of structural, chemical, or electrical abnormalities in the nervous system and are characterized by different signs and symptoms [160].

In 2006, the World Health Organization estimated that neurological disorders affect one billion people worldwide [121]. Neurological disorders are a major health concern because of their impact on both patients and their families. Most of them make affected people unable to carry out their normal daily activities. They also place a heavy burden on society because their diagnosis, management and treatment are very difficult and

expensive.

Deep knowledge, early diagnosis, and treatment of neurological disorders play a critical role in improving the life of affected patients. An increasingly used approach to support the study, diagnosis, and treatment of neurological disorders is to take advantage of the ever-increasing amount of digital medical data collected through specialized healthcare systems, sensors, and even consumer solutions such as mobile applications. In this thesis, we describe our work with two very different medical datasets that were created to study two different types of neurological disorders: disorders affecting motor control (e.g., Parkinson’s disease) and brain tumors. The two datasets are very different in terms of the type, quality, amount, and characteristics of the data, and they have been used for different purposes.

In the first part of the thesis, we present the process we followed and the results we obtained in the analysis of a large dataset of neuro-motor performance data collected in an unsupervised, non-clinical setting through a free and publicly available mobile application [167]. The app includes digital versions of standard motor tests that are used by neurologists in their classic pen-and-paper form to assess motor performance of the upper limbs in a clinical setting and identify symptoms of movement disorders such as Parkinson’s disease. The aim of the project was to collect data on a large scale and analyze the collected data to identify age-related behavior patterns of healthy subjects from different motor tests. These normative behavior patterns would then form the basis for early diagnosis of motor-related symptoms of disorders such as Parkinson’s disease. Because of the unsupervised data collection approach, we paid special attention to the data cleaning process, identifying various criteria and trying different solutions to remove incomplete and incorrect data. We then used a statistical approach to compare the motor performance of different age groups in the different motor tests with the goal of establishing whether the data revealed the degradation of human motor performance that is typical of aging [95, 149] and hence if the app, the considered motor tests,

and the measures we used to characterize motor performance are appropriate for the unsupervised collection and analysis of neuro-motor performance data. Subsequently, we explored using machine learning techniques to automatically classify users based on their motor performance. Since we were limited to use performance data of (presumably) healthy individuals, the classification problem was formulated as an age group identification problem, with the idea that a case that could not be clearly classified in its correct age group would be flagged for manual inspection from a neurologist. To this end, we first extracted additional motor performance measures from the data based on relevant literature and then used appropriate feature selection methods to identify the best features for the classification task. We then applied and evaluated the accuracy of two different machine learning techniques in a binary (2 age groups) and a multi-class (5 age groups) classification problem. Chapter 2 provides the background for the first part of the thesis introducing the topic of human motor performance assessment for motor disorders identification and presenting related work on the use of digital diagnostic tests for the evaluation of motor performance. The chapter includes a detailed description of MotorBrain, the mobile application that was used to collect the dataset analyzed in this part of the work, and concludes with a summary of our research goals. Chapter 3 gives an overview of the MotorBrain dataset in terms of size and user demographics, before focusing on the process followed to clean the data captured by the app based on multiple criteria. After presenting the performance measures we derived from the raw data, Chapter 4 describes two statistical analyses we did on the dataset, the first focusing on comparing performance of two age-groups that are often used in the literature, young adults and old adults, and the second being a more fine-grained analysis of the collected data in terms of a subdivision of the subjects in 5 age groups. Finally, Chapter 5 presents the machine learning process followed to automatically classify users by age group based on their motor performance.

The main topic of the second part of the thesis is radiomics based prediction of the

overall survival (OS) of patients suffering from High Grade Glioma (HGG), a major category of tumors affecting the brain, through the automatic analysis of brain images obtained with 3D Multi-parametric Magnetic Resonance Imaging (MRI). In this part, we used a standard dataset of MR images provided by the Brain Tumor Segmentation Challenge (BraTS). Our main focus was on two steps of the radiomic process: segmentation and feature selection. Radiomic features that are extracted from 3D Multi-parametric MRI and that are used to predict OS are sensitive to the variability in tumor subregions segmentation algorithms. While many algorithms have been proposed for automatic segmentation of brain tumor sub-regions, no evidence is available about how much these algorithms affect radiomic performance. Our work thus aimed to quantify the effect on OS prediction of variations in automatic segmentation of the brain tumor volume. In particular, we segmented tumor subregions using five state-of-the-art Deep Learning (DL) algorithms, creating 4 different types of segmentation models that we used to extract radiomic features. We also used the STAPLE label fusion method [142], to fuse the segmentation labels obtained from the five DL segmentation algorithms. We then evaluated the efficacy of the multi-region segmentation maps obtained from the individual DL algorithms and the STAPLE fusion method for the radiomics-based prediction of OS in HGGs. Finally, we evaluated the impact of stability analysis of radiomic features on OS prediction. To this purpose, we used the intra-class correlation coefficient (ICC) and overall correlation coefficient (OCCC) to quantitatively measure the robustness of radiomic features across seven state-of-the-art (independent) segmentation methods based on Convolved Neural Networks (CNNs). We employed two feature selection techniques to select discriminatory features: Minimum Redundancy, Maximum Relevance (MRMR) and Recursive Feature Elimination with SVM (RFE-SVM). Then we evaluated the effect of robust radiomic features for OS classification by incorporating stability into feature selection methods, considering both stable features (via ICC and OCCC) and discriminatory features (via MRMR and RFE-SVM).

Chapter 6 provides the background for this part of the thesis, introducing brain tumors and the task of OS prediction of HGGs with radiomics, describing the radiomics process, and presenting the state of the art on segmentation algorithms, feature robustness, and OS prediction. The chapter concludes with a summary of our research goals in this context. Chapter 7 presents the experimental methodology we adopted and the results we obtained in the evaluation of the impact of state-of-the-art segmentation algorithms on OS prediction with multiregional radiomics. Chapter 8 presents the experimental methodology we adopted and the results we obtained in the evaluation the robustness of radiomic features for OS prediction. Finally, Chapter 9 presents the conclusion of the thesis and discusses possible future work.

PART I
Computer-aided analysis of a large
neuro-motor dataset for upper limbs motor
performance assessment

2

Background: digital motor tests for upper limbs motor performance assessment

Neurologists use several different methods to evaluate human motor performance of the upper limbs to support the diagnosis of movement disorders such as Parkinson's disease. In recent years, mobile technologies have increasingly being used to this purpose due to their flexibility, familiarity, and capability to support the acquisition and analysis of motor performance data. MotorBrain [167] is one such mobile application that allows users to autonomously carry out different types of motor tests. The analysis of the dataset collected through MotorBrain is the focus of our work in this part of the thesis. In this chapter, after a brief introduction of the concept of motor control and of the related disorders, we survey digital diagnostic tests that have been proposed in the literature to assess upper limbs motor performance, focusing on mobile-based solutions.

We then introduce the MotorBrain application, describing the digital motor tests it includes and the types of data recorded by the app. We conclude by summarizing the scientific goals of our work with the MotorBrain dataset.

2.1 Motor control, movement disorders, and aging

Motor control is the process that allows people to perform motor tasks through the cooperation between the nervous system and the neuromuscular system [172]. People who suffer from medical conditions that affect their motor control have difficulty in controlling their bodies, such as regulating movement, stability, balance, coordination, and interaction with the outside world [144].

As people age, various physiological and anatomical changes occur in the brain that lead to a deterioration of their motor control. It is thus natural for motor performance to get worse with aging [95, 149]. At the same time, disorders that affect motor control are often strongly associated with age. As life expectancy increases, the number of people suffering from movement disorders will thus steadily increase. Kontis et al. [89], reported that the average life expectancy of people will be 85 years or more in 2030. In European countries such as Italy, Greece, Germany and Portugal, the number of elderly people is already higher than that of young people [126].

Among the many different neurological disorders that have an impact on motor control and that appear to have a relation with age, Parkinson's disease is certainly one of the most common and most studied, affecting approximately 8 to 18 out of 100,000 people each year. The prevalence of Parkinson's disease is 1% in adults aged 65 to 69 years and 1 to 3% in people older than 80 years [120]. Figure 2.1 shows the prevalence of Parkinson's disease in different countries around the world. It can be clearly seen that the percentage of Parkinson's disease in European countries is second only to China. Experts predict that the prevalence of brain diseases, including

Parkinson's, will quadruple by 2050, if not sooner.

Diagnosis and treatment of neurological disorders that affect motor control is difficult. The causes of these disorders are often unknown and diagnosis is typically based on motor symptoms. In the case of Parkinson's disease, there is some evidence that genetic factors, environmental factors, or a combination of both play a role [79]. However, no cure is known that slows the neurodegenerative process and treatment is mainly aimed at reducing the effects of symptoms. Motor symptoms advance aggressively and are more difficult to manage if they are not treated early and properly. It is thus crucial to diagnose the patient in the early stage of the disease, when symptoms are mild and often ignored, in order to make a proper prognosis.

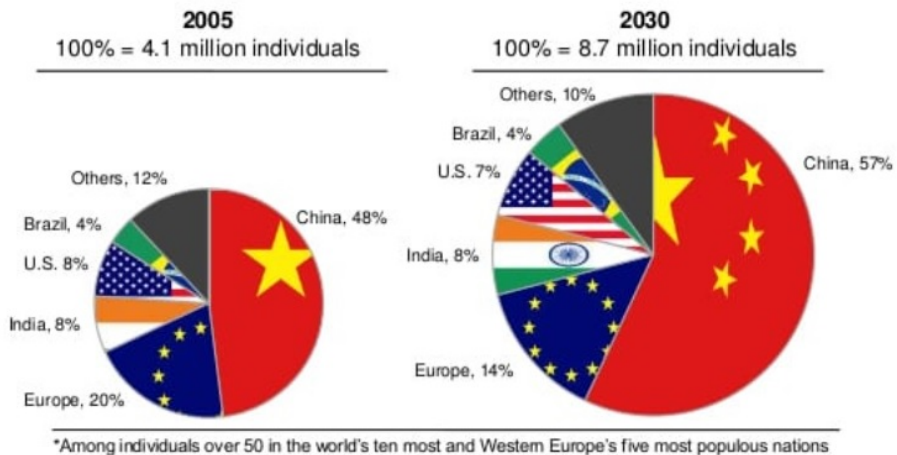


Figure 2.1: Distribution of individuals with Parkinson's disease by country from 2005 to 2030 [47]

2.2 Tests of motor control

There are no definitive tests to diagnose neurological disorders that affect motor control like Parkinson's disease (PD). Methods such as neuroimaging, laboratory tests, and

cerebrospinal fluid examinations are expensive and can only diagnose a disease in the later stages [92]. While four motor symptoms are typically considered as the classic identifying feature of PD (tremor, slowness of movement, rigidity, and postural instability), the signs and symptoms of movement disorders vary from patient to patient and show a progressive decline in both motor and cognitive abilities.

Some studies have reported that PD patients show more irregularities in upper limbs movements, such as abnormal hand movements, increased corrective movements, and decreased speed compared to healthy adults [49, 50, 42]. Over the years, neurologists have thus designed several noninvasive paper-and-pencil-based tests that can be used to assess upper limbs motor performance and, in general, to assess patients' motor skills or impairments. The handwriting (writing some text on paper) [117], finger tapping (tapping with the index finger on a specific surface) [4], trail making (drawing simple shapes on paper) [139], clock drawing (drawing a clock by writing numbers around the circle and then placing the clock hands on specific numbers) [60], and spiral drawing (drawing a spiral shape) [36] tests are all common examples. Digital versions of the paper-and-pencil tests have also been created and used to solve some of the limitations of paper-and-pencil based methods such as their unsuitability for repeated measurements.

However, these methods still have limitations. They do not capture critical performance changes over time, they usually provide small amounts of data on a limited number of people, and they are usually performed in a laboratory environment under supervision, making them not easily and equally accessible to everyone. In addition, when the tests are administered, subjects may already be showing signs of motor activity decline. If subjects already suffer from some motor dysfunction, they may not be able to adapt to new tasks and therefore may need some type of training sessions. Such training allows them to become familiar with the tasks, which has an effect on their performance [40].

2.3 Mobile apps for motor control testing

In recent years, mobile technology has improved to the point that many digital health applications for smartphones and tablets can be used to collect information about different parameters of the human body and support various informational, monitoring, or diagnostic tasks. Some apps use both mobile devices and wearable sensors and require a clinical setup to assess patients' symptoms. Other apps can be used autonomously by people to monitor their health. Examples of the different types of mobile health applications that help monitor the patient's condition include physiological health systems [154], health alerts with connection to medical professionals [78], calculators of the required insulin dose [72], and medical information systems [48].

Digital versions of the paper-and-pencil motor tests have been proposed in the form of mobile applications to overcome the previously mentioned limitations of the traditional versions. Many of these solutions have been developed to support the diagnosis and treatment of PD patients, allowing multiple and complex PD symptoms to be captured.

Lauraitis et al. [93] proposed a model for the digital screening of neurological impairments. They collected a dataset about 15 subjects (8 healthy and 7 with neurological disorders) that used a mobile app to carry out 16 tasks, of which 12 are a digital version of the Self-Administered Gerocognitive Examination (SAGE), and the others are finger tapping and speech recording tasks. From the dataset, they extracted a set of 238 features. The discriminatory features were selected using a variety of methods, such as principal component analysis, wrapper subset evaluation, and classifier attribute evaluation. A subset of the features was used to train 13 different machine learning classifiers. The final results were obtained using the average of the probabilities and showed 96.12% accuracy in classifying neurologically impaired and healthy individuals.

Creagh et al. [39] proposed a smartphone-based Draw a Shape (DaS) test to assess upper limb function in people with multiple sclerosis. They collected data from 93

subjects (22 healthy subjects and 71 multiple sclerosis patients). Each participant was required to draw various shapes, such as lines, circles, squares, spirals, and 8 figures, on a cell phone with their dominant and non-dominant hand. Temporal and spatial features were extracted from the shape drawings. The dimensionality of the features was reduced using the LASSO method. Random Forest and Support Vector Regressors were used to evaluate the relationship between the DaS tests and the associated 9HPT which is used to measure dexterity in patients with various neurological disorders. The study found that shapes drawn with the nondominant hand more accurately predicted 9HPT time than those drawn with the dominant hand.

Iakovakis et al. [74] analysed key dynamics during typing on a mobile touchscreen in a clinical setting to measure fine motor skills performance. They recorded the typing of 18 PD and 15 healthy subjects, all of whom were right-handed and over 40 years of age. Keystroke dynamics variables such as Hold Time (HT), Flight Time (FT), and Normalised Pressure (NP) were extracted from the typing sessions. A logistic regression classifier was trained with statistical features (mean, standard deviation, kurtosis, skewness, and covariance) extracted from the keystroke dynamic variables. The classification task was performed in two stages. In the first stage, three models, one for each dynamic variable, were trained and used for prediction. In the second stage, the results of the three models were combined and fed into a classifier to obtain a final prediction of whether the typing session belonged to a healthy or PD subject.

In a study by Zham et al. [183], 31 subjects with Parkinson’s disease (PD) and 31 healthy subjects each performed spiral drawing and handwriting tasks on a tablet using a digital pen. 14 spatial and temporal features were extracted. The top five features were selected using the relief F method. A Naive Bayes classifier with spiral drawing features was used to classify the PD and healthy subjects. The results showed an AUC of 0.933 when classifying PD and healthy subjects. Statistical analysis of features extracted from the subjects’ drawing and handwriting tasks revealed a significant difference between

the two groups.

Arroyo Gallego et al. [5] quantified motor skill impairment using touchscreen typing and alternate finger tapping tests. They collected data from 24 PD and 27 healthy subjects. Statistical features such as kurtosis, skewness, and covariance from touch screen typing and the average number of finger taps from dominant and non-dominant hands tapping were extracted. Feature selection was performed with an L1 regularizer and used to train a linear support vector machine classifier. The study yielded an Area Under the Curve (AUC) of 0.91 for the best feature typing sessions and an AUC of 0.85 for the finger tapping tests when classifying PD and healthy subjects.

Wissel et al. [116] used a mobile application called iMotor to distinguish motor functions of PD subjects from those of healthy people. iMotor includes finger tapping, pronation-supination (based on touching the screen alternatively with the palmar and dorsal surface of hand), and reaction time neurological tasks. 19 PD and 17 healthy subjects participated in the study. Variables for analysis were derived from recorded screen pixels (x, y) and included total number of taps, tap accuracy, tap speed, tap interval, tap duration, and reaction time. The results of a multivariate logistic regression analysis for the extracted features show that reaction time was the best predictive variable with an AUC of 0.90 for PD.

ParkNosis [145] is an Android-based smartphone app that provides users with hand tremors, spiral drawing, and tapping tests, as well as a questionnaire to assess their motor skills. The app was used by the authors to collect data from 11 participants (PD and healthy subjects) belonging to different age groups, the majority of whom were right-handed. The analysis was performed on the mobile phone. Based on the analysis results, the authors defined a scale to evaluate the patients and inform them about their PD status.

2.4 The MotorBrain app

MotorBrain [167] is a mobile application that includes digital versions of four trail making and two finger-tapping paper-and-pencil motor tests. It was designed and developed by a research group at the HCI Lab of the University of Udine and was released in Italy as a free app for major mobile platforms (iOS, Android, Windows Phone). Unlike most of the mobile solutions mentioned in the previous section, MotorBrain was not meant to be used exclusively in a clinical setting or for a specific medical condition, albeit it could be used as such if needed. It was created as a data collection and assessment tool for the millions of smartphones and tablets available to the general public. Its interface was designed to be easy to use for individuals of any age, letting them interact naturally and directly with the tests using their fingers. The app provides a preliminary assessment of motor performance directly on the device but it also sends detailed motor test data to a remote server for storage and further detailed analysis. Overall, more than 10000 users downloaded and used the app since its release.

2.4.1 Motor tests

The motor tests included in MotorBrain are organized in three groups consisting of two tests each. The groups are defined based on the primary characteristic of motor control they focus on (see Figure 2.2): two trail making tests have been designed to primarily measure accuracy (how accurate users are in following the path displayed in the test), the other two trail making tests focus on speed (how fast users are in following the path displayed in the test), while the two finger tapping tests measure reaction time (how quick users are to tap targets appearing in the test). Accuracy, speed, and reaction time are common measures of users' motor skills that have been employed in other studies in the literature [18, 71, 138]. Only one category is enabled when the user launches the app for the first time. The user must complete all tests in that category before the next

one becomes available. To complete a test, users must perform three repetitions of it, first with their dominant hand and then with their non-dominant hand.

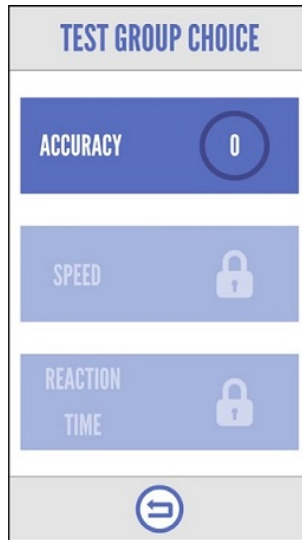


Figure 2.2: MotorBrain tests are grouped in three categories based on the motor control measure they focus on. The Accuracy category is enabled in this example. The number on the right shows the number of times a users completed all the tests within that category. [167].

Accuracy tests

The two tests in the accuracy group, called Circle-A and Square, require users to follow a path displayed on the screen as accurately as possible and differ for the shape of the path to follow.

The Circle-A test requires users to move their index finger on a colored ring (see Figure 2.3A), following its entire length once as accurately as possible and without lifting their finger. The movement must be performed clockwise when users are using their right hand and counterclockwise when they are using their left hand. The starting point is displaced by 30 degrees on the left (right) for clockwise (counterclockwise) movements. The diameter of the outer circle is 4cms while the thickness of the path is

0.5cms. A repetition ends when the total distance moved by users with their finger is equal to the circumference of the ring (in its middle point), when users lift their finger off the screen for more than 0.15 seconds, or when 10 seconds have passed from the start of movement.

In the Square test, users have to move their index finger on a square frame (see Figure 2.3B), following its entire length once as accurately as possible and without lifting their finger. As in Circle-A, movement must be clockwise when users are using their right hand and counterclockwise when they are using their left hand. The starting point is the top left vertex of the frame for clockwise movement and the top right vertex for counterclockwise movement. The side of the square frame is 4cms and its thickness is 0.5cms. A repetition ends when the total distance moved by users with their finger is equal to the perimeter of the square frame (in its middle point), when users lift their finger off the screen for more than 0.15 seconds, or when 10 seconds have passed from the start of movement.

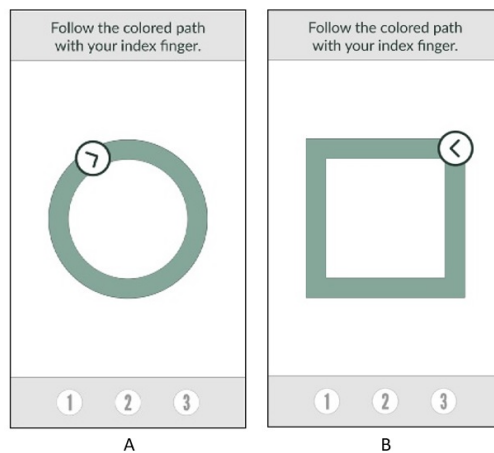


Figure 2.3: MotorBrain tests designed to measure accuracy: (A) Circle-A: follow the colored ring starting from the highlighted position. (B) Square: follow the square frame starting from the highlighted position.

Speed tests

The two tests in the speed group, called Circle-S and Path, require users to follow a path displayed on the screen as fast as possible and differ for the shape of the path to follow and the number of times the path can be repeated.

In the Circle-S test, users have to move their index finger on a colored ring (see Figure 2.4A), following its entire length as many times as possible and without lifting their finger, within a 7 seconds time limit. The ring has the same diameter as in Circle-A but its thickness is slightly larger (0.7cms). As in the accuracy tests, movement must be clockwise when users are using their right hand and counterclockwise when they are using their left hand, and the starting point is displaced by 30degrees . A repetition ends when the 7 seconds time limit is reached or when the user lifts her finger for more than 0.15 seconds.

The Path test requires users to move their index finger over a path comprised of four interconnected lines (see Figure 2.4A), following its entire length once as fast as possible and without lifting their finger, within a 5 seconds time limit. Each line is 3.72cms long with a 0.6cms thickness and the angle between each pair of connected lines is 19.80 degrees. When doing the test with their right hand, users start at the top left position and move to the bottom of the path. When using their left hand, the path is mirrored and users move from the top right position to the bottom of the path. A repetition ends when the 5 seconds time limit is reached, when the user raises her finger for more than 0.15 seconds or when the total distance moved by users with their finger is equal to the length of the path.

Reaction time tests

The two tests in the reaction time group, called Tapping-2 and Tapping-4, require users to tap as fast as possible on a target that appears on the screen and differ for the number

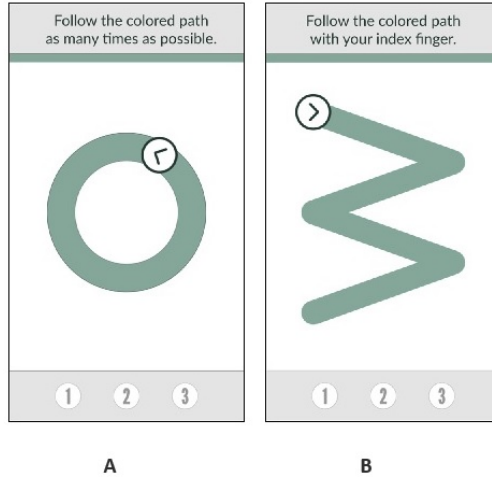


Figure 2.4: MotorBrain tests designed to measure Speed: (A) Circle-S: follow the ring as many times as possible within a time limit. (B) Path: follow the path once as fast as possible.

of possible positions where the target could appear.

In the Tapping-2 test, the target to tap can appear alternatively in two different but aligned positions. When the target is active, it is displayed as a colored circular button of 1.6cms diameter with a viewfinder consisting of three concentric circles (see Figure 2.5A). When the target is disabled, it is displayed as a light grey area without viewfinder. When users tap on an active target, the target deactivates at that position and becomes active at the other position. Users have to tap on the active target as many times as possible during a 10-second interval, starting when the user taps on the first target. At the beginning of the test, the active target is on the left for a test done with the right hand and on the right for the left hand.

In the Tapping-4 test, the target is of the same type as in Tapping-2 but can appear in 4 different positions (see Figure 2.5B). Unlike in Tapping-2, the initial position of the active target as well as the position where the active target will appear after it has been tapped once in its current position are randomly selected. As in Tapping-2, users have to tap on the active target as many times as possible during a 10-second interval.

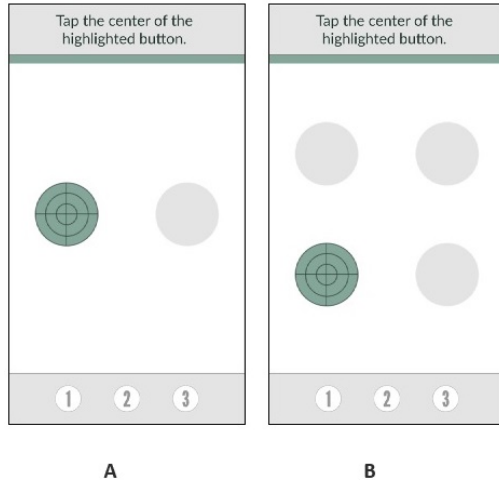


Figure 2.5: MotorBrain tests designed to measure Reaction Time: (A) Tapping2: tap as many times as possible on the active target, which is visible in one out of two possible positions and changes position after each tap. (B) Tapping4: tap as many times as possible on the active target, which is visible in one out of four possible positions and changes position after each tap.

2.4.2 Recorded data

MotorBrain automatically records four major categories of data: (a) data related to the screen touch events that are generated by users when they carry out motor tests, (b) demographic data about users, (c) technical data about the mobile device running the app, and (d) derived data about user’s performance in the motor tests.

Screen touch event data

The bulk of the data recorded by MotorBrain consists of the coordinates (x, y) in pixels and the corresponding timestamps t_n of the events generated each time the user touches the screen or changes finger position on the screen during one of the motor tests. The temporal sequence of coordinates allows us to recreate the trajectories followed by users during the trail making motor tests and the sequence of points tapped by users during the finger tapping motor tests, and compare them to the expected outcome based on the underlying stimulus (path to follow or target to tap).

Screen touch event data is organized hierarchically. The set of touch event data related to one repetition of a motor test is an individual record with its own identifier. Each complete run of a motor test consists of 6 records, 3 repetitions each for the dominant and non-dominant hand. At a higher level, motor tests are organized in sessions which are created when a user logs into the app, with each session associated to a specific user and having its own timestamp.

User data

User data recorded by the app consists of the age, gender, and dominant hand of each user, which are asked through a demographic questionnaire when users run the app for the first time. Users can also choose a nickname that will be locally associated to their profile on the device. The app also records answers to yes/no questions that are presented to users at the beginning of each session of use, about whether they suffered at the time of the session from headache or migraine, or any illnesses or pain at the wrist, elbow, or hand, and whether they were under the effects of drugs or abused substances that could impair performance.

Device data

Device data recorded by the app consists of the device model (as communicated by the device itself on request of the app), screen height, screen width, and screen resolution in terms of dots per inch. Since the last three measures cannot always be obtained from the device, the app includes a fallback mechanism that tries to derive the data at run-time from an external database.

Derived performance data

To provide users with an assessment of their motor performance, the app computes three measures that are shown at the end of each motor test and that correspond to

the three primary characteristic of motor control that each of the three groups of motor tests were meant to measure, i.e., accuracy, speed, and reaction time:

- Accuracy is computed as the ratio between the distance covered by the index finger on the ring (Circle-A) or the frame (Square) and the optimal distance (circumference of the ring or perimeter of the square at their middle point), expressed as a percentage.
- Speed is the ratio (in cm/s) between the distance covered on the ring (Circle-S) and on the path (Path) and the total time spent in performing the movements.
- Reaction time is the mean of all the times (in seconds) elapsed between the appearance of a target button on the screen and user's tapping on that target, excluding the first tap.

In addition to being computed and displayed to users at run-time, these measures are also recorded for analysis.

2.5 Our research goals

The analysis of the dataset collected through MotorBrain is the subject of the first part of this thesis. A dataset of this type could be used in principle to build a normative reference of motor performance, which is currently unavailable in neurology and could be effectively used for different purposes. It could allow neurologists to investigate the physiological aspects that are involved in the aging of the population's neuromuscular system [33]. It could make it possible for neurologists to assess the evolution of a movement disorder over time, to study how the motor skills of individuals change over time during the rehabilitation process of a movement disorder, and to assess the effect of pharmacological or physical therapies. It could offer neurologists the possibility to carry out an early differential diagnosis of movement disorders, e.g. Parkinson's disease,

by comparing the motor performance data of a specific patient with normative data for the same age group. As previously mentioned, early differential diagnosis is very important because it allows neurologists to start clinical treatments in a timely fashion, thus reducing the negative effects that degenerative disorders have on patients.

All these goals can be potentially reached if there is sufficient evidence that the collected dataset is meaningful in the assessment of users' motor performance. In particular, if it characterizes the normal motor behavior of the population and how it changes with aging. The analysis we performed on the dataset was thus aimed at discovering if the data was consistent with the expected motor performance patterns of healthy subjects, and especially with the degradation of human motor performance that is typical of aging [95, 149]. We also developed machine learning models that can be used to classify performance results based on the age-based normative behavior identified in the analysis and thus help neurologist to identify neurological disorders at an early stage by automatically comparing new data with the available normative data.

3

The MotorBrain dataset: overview and data cleaning process

The MotorBrain app was designed to collect data remotely from users without any external supervision by experts who could guide users in carrying out the motor tests. Even if the app includes simple textual and visual instructions, there is no guarantee that users did not do the tests incorrectly, either voluntarily or because they did not understand the instructions. In addition, technical issues related to the device ecosystem, data transmission, and data storage could also affect the quality of the collected data. This is why it is important to perform a data cleaning process before being able to use the data for further analysis. In this chapter, we first provide an overview of the MotorBrain dataset in terms of user demographics. We then describe the different criteria we used to clean the data, removing incomplete, inconsistent and inaccurate

records, and the results of the process.

3.1 Overview of the MotorBrain dataset

Overall, we received data from 12503 users. Since the app was available in Italian only and was distributed in the Italian version of the app stores, all users are supposed to be Italian. Demographic information about the users is summarized in Table 3.1.

Table 3.1: Demographics of MotorBrain users in the collected dataset.

| No. of Users | |
|-----------------------------|---------|
| Total | 12503 |
| Male | 7100 |
| Female | 5403 |
| Age of users (years) | |
| Range | 6 – 110 |
| Mean | 26 |
| Dominant hand | |
| Right | 11321 |
| Left | 1182 |

The age of users ranges from 6 to 110 years with a mean of 26, showing that the app attracted users from all age groups, including older adults, but that a large percentage of users were younger, which is consistent with the typical demographic distribution of mobile app users. This can be clearly seen in the histogram in Figure 3.1 which shows the age distribution of users. If we group users in 20 years buckets, we have 5032 users

in the 1 – 20 years range, 5535 in the 21 – 40 years range, 1656 users in the 41 – 60 years range, 250 users in the 61 – 80 years range, and 30 users above 80 years.

As shown in Table 3.1, the number of male users is slightly higher than the number of female users. This also appears to be consistent with studies on the use of mobile health apps [87]. The number of users who declared that the right hand was their dominant hand (11321) is largely higher than the number of users who declared the left hand as their dominant hand (1182), which is in agreement with the prevalence of right-handedness in humans [133].

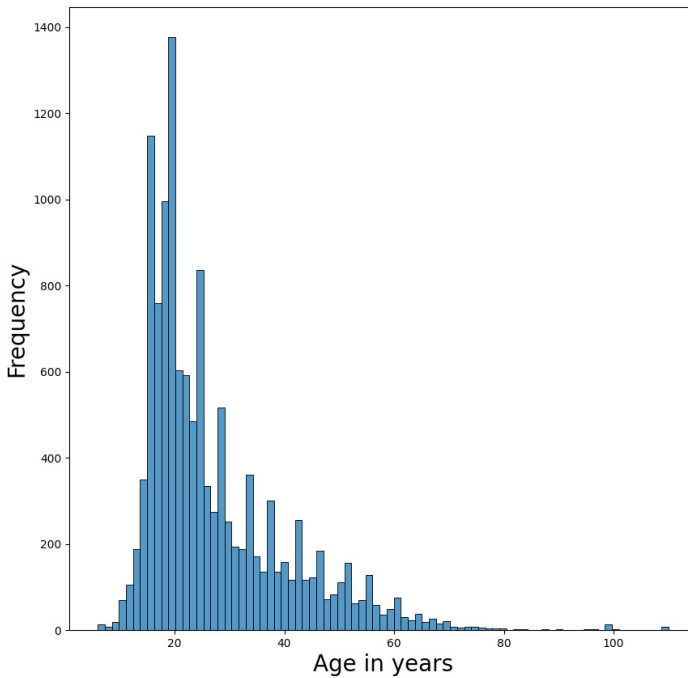


Figure 3.1: Age distribution of users

If we consider the number of users who carried out the individual motor tests at

least once, we obtain 12185 users for Circle-A, 11713 users for Square, 11000 users for Path, 10300 users for Circle-S, 9453 users for Tapping-2, and 8194 users for Tapping-4. There is a noticeable decline in the number of users as we move from accuracy tests to speed tests to reaction time tests, with $1/3$ of the users who completed Circle-A not even reaching the Tapping-4 test. This is most likely a consequence of the fact that only the accuracy tests are available when the user launches the app for the first time and the other two categories unlock only when the user completes all tests in the previous category. This design choice was made to have all users perform the motor tests in the same order so that their initial motor performance in each test could be comparable in terms of previous experience. Considering that most mobile app users interact with apps in short time bursts and repeated usage is not common for most apps, users who did not have the time, patience or interest to unlock all tests in one session may account for the above effect.

3.2 The data cleaning process

The basic unit of interest for the data cleaning process is the individual record containing touch event data for one repetition of a motor test. The different steps of the process led to the removal of records based on criteria related to the quality of each record itself, to the properties of other records in the same run of a motor test (consisting of 3 repetitions), to the user condition, and to the device characteristics. Overall, the initial number of records in the dataset was 88108 for Circle-A, 82244 for Square, 73521 for Path, 69143 for Circle-S, 62541 for Tapping-2, and 55230 for Tapping-4. These numbers show that, on average, users did about 7 repetitions for each motor test, which is just slightly higher than the 6 repetitions needed to complete a test with both hands.

The process consisted of 5 record removal activities, each one related to a different type of issue with the data, the device, or the user: (a) small or erroneous screen size,

(b) missing values, (c) incomplete runs, (d) anomalous psychophysical state of the user, and (e) incorrect/incomplete performance. These types of issue have been identified by directly looking at samples of the data or by analyzing visualizations of the data. While the removal activities were carried out in a specific order, they are independent of each other and could have been completed in any order.

3.2.1 Small or erroneous screen size

A significant amount of records was removed from the dataset because the screen of the device was too small or was not recognized correctly.

MotorBrain uses information about device screen size and screen density (in dots per inch) to determine how to properly display graphical elements in tests. For each motor test, the application scales the graphical elements of the test to make sure they are displayed with the same physical size, regardless of the specific model of mobile device used. Such information is obtained directly from the device and is stored together with test data in the recorded dataset.

To draw the graphical elements with the sizes reported in Chapter 3 and make user's performance comparable across different devices, the diagonal display size of the device should be 4 inches or more (most smartphones and tablets in the last few years meet this requirement). On devices with a screen size smaller than 4 inches, the app reduces the size of graphical elements because the space available is insufficient. To properly support comparisons of the motor tests results, we thus removed all records associated to devices with small screen size (less than 4 inches).

Another issue appears when the screen information obtained from the device is incorrect. In such case, the application displays targets with an incorrect size, making all tests with those targets unusable for subsequent analysis. By manually comparing the device data saved in the dataset with the technical specifications of the corresponding devices available from device manufacturers, we found all instances of incorrectly rec-

ognized screen sizes and screen densities. We then removed from the dataset all those records associated to devices for which the difference between stored and actual device information (in terms of derived screen size) was larger than ± 0.075 inches. The threshold was defined in collaboration with the neurologists who participated in the design of MotorBrain.

These screen size issues led to the removal of 22318 records for Circle-A, 20365 records for Square, 20235 records for Path, 16183 records for Circle-S, 14938 records for Tapping-2, and 14148 records for Tapping-4, including almost all data collected from Apple iPad users.

3.2.2 Missing values

Missing values in the MotorBrain dataset are all those data values which are encoded as nulls and can belong to any of the data categories that are collected by the application (device data, user data, etc.). Null values are stored in the database when the received data is not complete, typically because of technical failures related to data generation, transmission and storage (most commonly, networking problems).

In the literature, different techniques have been proposed to deal with missing values and these typically depend on the type of missing data mechanism. If the causes of the missing data are unrelated to the data and occur entirely at random, the data are said to be missing completely at random (MCAR). If the missingness is not random but can be fully accounted for by variables where there is complete information, then the data are missing at random (MAR). If neither MCAR nor MAR holds, then we speak of missing not at random (MNAR).

The only possible events that led to missing values in our case are random technical failures that are independent of the variables of interest, the approach we used to deal with the missing data was to remove the records associated to the missing data. This led to the removal of 1325 records for Circle-A, 1180 records for Square, 1159 records for

Path, 1035 records for Circle-S, 582 records for Tapping-2, and 452 records for Tapping-4.

3.2.3 Incomplete runs

Each test required users to repeat the same task three times with their dominant hand and then three times with their non-dominant hand. Some users performed the tests only 1 or 2 times rather than 3, probably because they had to stop interacting with the app for external factors (e.g., calls, notifications, interruptions, etc.) or because they got bored with the tests. To keep consistency when comparing user performance, we only considered full runs of 3 repetitions, removing all records related to partial runs from the dataset. This led to the removal of 84 records for Circle-A, 81 records for Square, 43 records for Path, 77 records for Circle-S, 110 records for Tapping-2, and 89 records for Tapping-4.

3.2.4 Anomalous psychophysical state of the user

As previously mentioned, the MotorBrain app asks users some questions to assess their psychophysical state before starting a test session. The questions are meant to determine if users suffer from conditions (headache, pain, drug or alcohol alteration) that would affect their motor performance. We removed from the dataset all records of users who answered positively to one or more of these questions. This led to the removal of 2097 records for Circle-A, 1947 records for Square, 1698 records for Path, 1578 records for Circle-S, 905 records for Tapping-2, and 1245 records for Tapping-4.

3.2.5 Incorrect/incomplete performance

The final and most complex step in the cleaning process concerned all those cases where the user did not properly follow the instructions of each test, leading to incorrect and/or

incomplete results that are not considered representative of normal motor performance. Possible motivations for these cases are that users did not read the instructions correctly (misinterpreting what they had to do) or did voluntarily perform tests without following the instructions. To identify these cases, we defined a set of exclusion criteria that differ from test to test. The criteria were defined in collaboration with the neurologists who co-designed the MotorBrain app. Figure 3.2 shows examples of complete/correct and incomplete/incorrect performance for each test.

In the **Circle-A** test, a complete/correct result would correspond to a full circle. To identify incomplete/incorrect results, we defined two exclusion criteria:

1. Divide the drawn trajectories based on four 90 degrees slices. The length of the trajectory in each slice must be greater than zero. This loose constraint identifies cases in which the user stopped the test before reaching the end of the circle.
2. Calculate the centroid of the shapes drawn by users:

$$C_x, C_y = \frac{\sum_{i=1}^N x_i}{N}, \frac{\sum_{i=1}^N y_i}{N} \quad (3.1)$$

where (x, y) are the coordinates of the points of the shape as recorded by the app. Use the adjusted box plot method to find minimum and maximum thresholds for C_x and C_y . Any shape whose C_x or C_y are outside the bounds of these values is considered incorrect.

For the **Square and Path** tests, a complete/correct result would correspond to four interconnected lines. We considered the four lines separately. We then defined two exclusion criteria based on the length of these lines (simplified as the crow-line distance between the first and last point of each line):

1. The length of each line must be greater than zero. As in the Circle-A test, this is a loose constraint to identify an incomplete trajectory.

2. Compute the average line length for each repetition. Use the adjusted box-plot method on these values to find the lower bound on length, outside of which a value would be considered an outlier and identify an incorrect line.

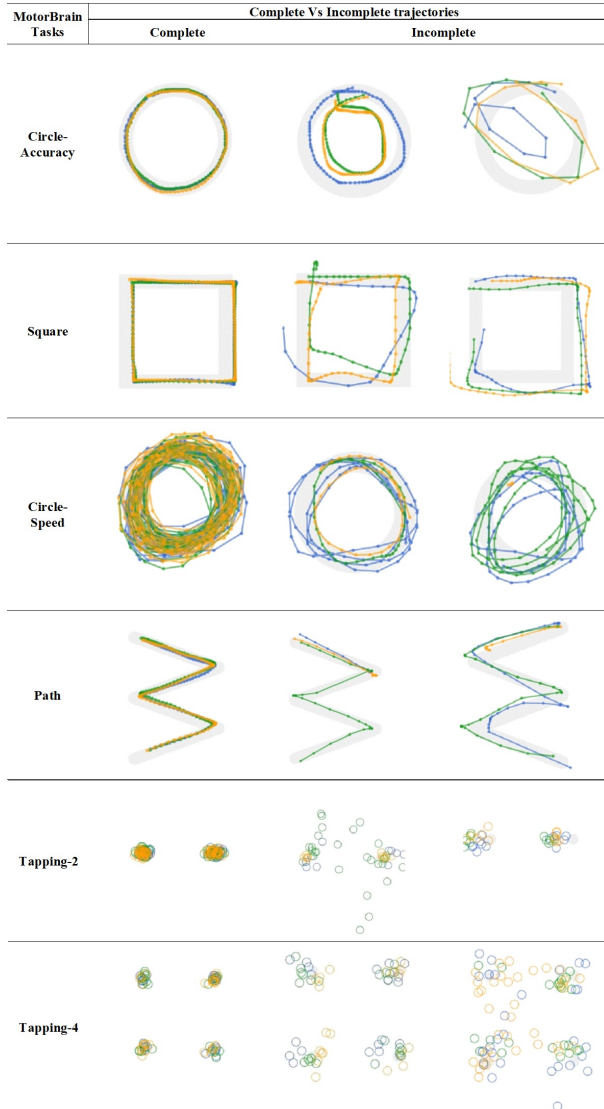


Figure 3.2: Examples of complete/correct vs incomplete/incorrect performance in the six motor tests. Each image shows all 3 repetitions of a test run with a different color.

In the **Circle-S** test, users were asked to draw as many circles as possible within the allotted time. In this case, the two exclusion criteria we considered were:

1. The total movement time must not be lower than 7 seconds. This identifies cases in which the user stopped the test before time.
2. Use the adjusted box-plot method on the number of circles drawn by users to find the lower bound on this number, outside of which a value would be considered an outlier and identify an incorrect test. This identifies cases in which the user completed too few circles with her finger in the allotted time.

For the **Tapping-2** and **Tapping-4** tests, we considered only one criterion based on the number of taps:

1. The total number of taps outside the three concentric rings of a target must be smaller than the number of taps inside the rings.

Overall, the application of these different criteria led to the removal of 2272 records for Circle-A, 1686 records for Square, 1008 records for Path, 26976 records for Circle-S, 49 records for Tapping-2, and 79 records for Tapping-4. The high number of records removed for Circle-S correspond to cases in which users did complete just one circle like in Circle-A, probably because they misunderstood the instructions of the test.

3.3 Results

Figure 3.3 shows a flow diagram of the full data cleaning process with the results of each cleaning activity and the final number of records. After the process, the number of records left was 60012 for Circle-A, 56985 for Square, 23294 for Circle-S, 49378 for Path, 45957 for Tapping-2, and 39217 for Tapping-4.

Overall, this shows the heterogeneity of issues that can affect the quality of health data collected without expert supervision through a mobile application, even when the

application is carefully designed to be as easy to use and self-explanatory as possible. A proper multi-step data cleaning process is thus essential to obtain a final dataset that could be used for analysis.

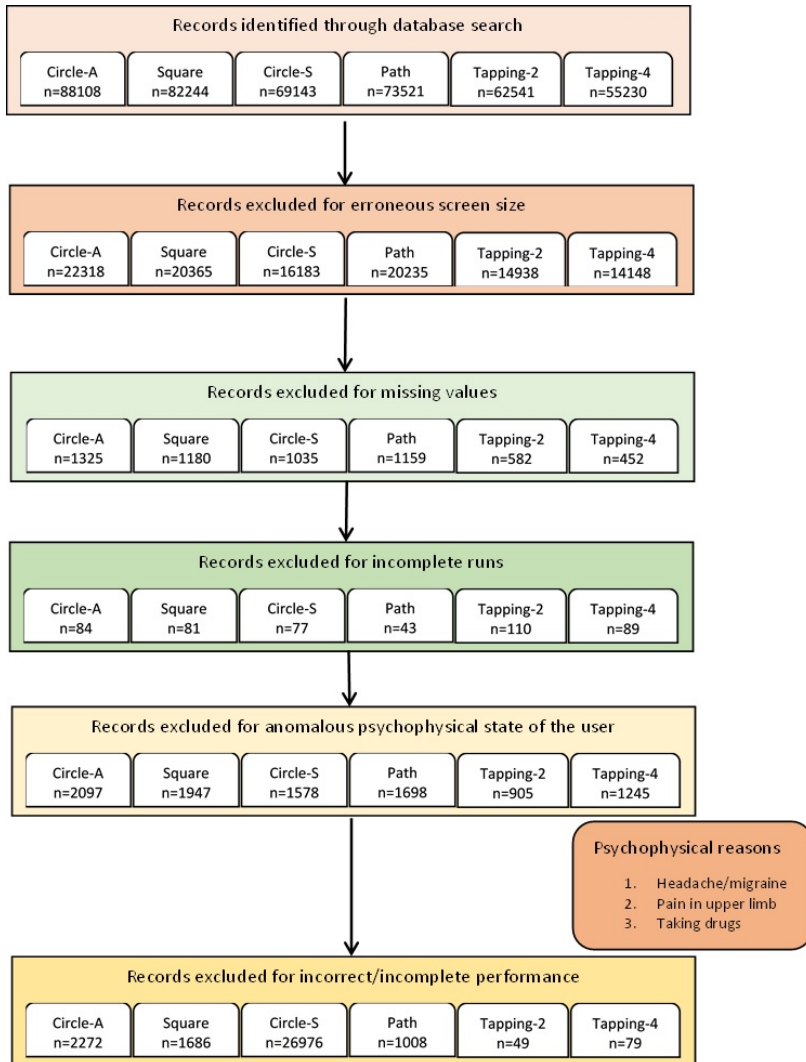


Figure 3.3: Flow diagram of the cleaning process showing the results of each activity in terms of the number n of records removed in that step.

4

Statistical analysis of the MotorBrain dataset

Once we had a cleaned dataset to work with, we needed to understand whether user performance as could be measured from the available data revealed expected patterns such as the typical degradation which is common with aging [95, 149] and the marked performance asymmetry between dominant and non-dominant hand [163]. This would provide evidence that MotorBrain allows to collect meaningful data in an unsupervised non-clinical setting, and could be used by neurologists to investigate normative behavior patterns and support the diagnosis of motor impairments.

The approach we followed to investigate the characteristics of users' motor performance is based on statistical analysis and its results are presented in this chapter. We first describe in detail the set of performance measures we derived for the motor tests included in MotorBrain to support the detailed analysis of the collected data. Then, we present the results of the two analyses we performed on the data. The first analysis focused on users belonging to two specific groups of users: young adults in the 18-30

years range and old adults in the 50-75 years range. The second analysis was more extensive and more fine-grained, including all users over 18 years old, divided in 5 age groups.

4.1 Performance measures

MotorBrain provides users with an assessment of their motor performance based on three primary measures (accuracy, speed, and reaction time) that characterize the three groups of motor tests and that are computed internally as described in Chapter ???. While these are important measures, they are not sufficient to fully capture the spatial and temporal aspects of the tasks to better characterize the behavior of users. Before performing the statistical analyses, we thus derived a more comprehensive set of measures for each of the two main types of motor tests in MotorBrain (trail making and finger tapping). The code for measure extraction was implemented in Python 3.6 and makes use of the points (x, y) touched by users on the screen and the corresponding time stamps t_n .

4.1.1 Trail making measures

Four of the motor tests included in MotorBrain (Circle-A, Square, Path, Circle-S) are trail making tests that require users to draw or follow a path on the screen. To better capture the motor performance of users in these tests, we computed three measures: error, speed and movement smoothness.

Error

Error measures the deviation of the user trajectory from the expected template path displayed on the screen. It is related to the accuracy measure provided by MotorBrain. It is computed in a different way for motor tasks based on circular paths (Circle-A and Circle-S) and for motor tasks based on linear paths (Square and Path).

For the **Circle-A** and **Circle-S** motor tests, which are based on a circular path, we first computed the average distance of each point of the trajectory followed by a user from the center of the circular target path. The distance between a point touched by the user (x_i, y_i) and the center of the target shape (x_c, y_c) is calculated as

$$r_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \quad (4.1)$$

The error is then computed using the following formula

$$Error = \sum_{i=1}^n \frac{(r_i - R)^2}{n} \quad (4.2)$$

where R is the template radius (distance between the (x, y) points of ideal trajectory to the center of the shape) and r_i is the radius calculated from the user drawn trajectory. If a user perfectly follows the required trajectory, the error would be zero.

For the **Square** and **Path** tests, the error is based on the length of the trajectory followed by the user, based on the following formula

$$Error = (covered_distance - template_length)/template_length \quad (4.3)$$

Here *covered_distance* is the total distance covered by the user while performing the test and *template_length* is the perimeter of the template path in the case of the Square test and the sum of the length of the four connected template lines for the Path test.

Speed

Speed in a specific test is measured by considering the distance covered in the unit of time and is calculated using the following formula

$$speed = \frac{\sum_{i=1}^N \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}}{\sum_{i=1}^N (t_{i+1} - t_i)} \quad (4.4)$$

where (x_i, y_i) and (x_{i+1}, y_{i+1}) are two consecutive points in the trajectory drawn by the user and $t_{i+1} - t_i$ is the time taken by the user to move from one point to the next. Speed is measured in *cm/sec*.

Movement smoothness

Smoothness is a measure of well-coordinated muscle movements, smooth coordination of wrist and fingers. The smoothness of the discrete arm movements in accuracy and speed derived tasks is an important characteristic of individuals healthy and well-trained motor behavior. The literature shows that log dimensionless jerk is a valid measure to quantify movement smoothness, especially for upper limb movements [65]. The log dimensionless jerk can be measured with the following formula:

$$LDLJ = -\ln |DLJ| \quad (4.5)$$

with

$$DLJ = -\frac{(t_2 - t_1)^3}{v_p^2} \int_{t_1}^{t_2} \left| \frac{d^2v(t)}{dt^2} \right| dt \quad (4.6)$$

where $v(t)$ is the movement speed, t_1 and t_2 are the movement start and end times, v_p is the maximum velocity between time t_1 and t_2 .

4.1.2 Finger tapping measures

For tapping tests, the important measures that allow to describe the age-related effects of human motor performance are the reaction time and the tap precision of each tap on the screen.

Reaction time

Reaction time, which captures the temporal aspect of a finger tapping task, is measured by computing for each target i the difference between tap time tt_i , i.e. the time when the user tapped on the active target after it appeared, and stimulus time st_i , i.e. the time when target button i became active

$$R_i = (st_i - tt_i) \quad (4.7)$$

and then taking the median value as a statistical measure of central tendency. Median is used in studies about finger tapping tasks as a preferable measure over the mean because it is a more robust measure for skewed data [74, 71].

Error

Error, which captures the spatial aspect of a finger tapping task, is measured by computing the distance between the coordinates (x_i, y_i) of the user's tap on the screen and the coordinates (x_c, y_c) of the center of the corresponding target, using the usual distance formula

$$d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \quad (4.8)$$

and then taking the median value, as in the case of reaction time.

4.2 Young adults vs. old adults data analysis

Before MotorBrain was made publicly available in the stores, it was used in a controlled lab-based study [167] to compare the motor performance of two groups of healthy subjects, young adults in the 18 – 30 years range and old adults in the 50 – 75 years range. The study involved 133 subjects and focused on the main characteristics of motor control

(accuracy, speed and reaction time) computed by the application. Results showed that young adults performed the trail making tests more accurately and quickly compared to old adults and reacted more quickly in the tapping tests.

We focused our first analysis of the MotorBrain dataset on the same two groups of users who were considered in the previous controlled study so that we could compare the results. However, we also analyzed the effect of handedness, which was not explicitly considered in the previous study, in addition to the effect of age. The number of subjects involved in the analysis in the 18 – 30 and 50 – 75 age groups differs by motor test and is reported in Table 4.1. These differences are a consequence of the different number of users who completed the different motor tests and of the cleaning process that we did on the data as described in Chapter 3.

Table 4.1: Number of users by group for the different motor tests.

| MotorBrain Tests | Young (18 - 30) | Old (50 - 75) |
|-------------------------|----------------------------|--------------------------|
| Circle-A | 3717 | 420 |
| Square | 3491 | 403 |
| Circle-S | 1093 | 249 |
| Path | 3420 | 417 |
| Tapping-2 | 2970 | 378 |
| Tapping-4 | 2573 | 358 |

After extracting the data about the two age groups from the dataset, we performed an additional filtering step on it, keeping only the first run for each motor test, i.e., the data about the first 3 repetitions of a test with each hand. This was done because a user could do multiple runs of each test if she wanted but we only focused on first-time performance to keep the comparison consistent across users.

The analysis focused on the performance measures presented in Chapter 3: error, speed and movement smoothness for trail making tests (Circle-A, Square, Path, Circle-

S), and reaction time and error for finger tapping tests (Tapping-2 and Tapping-4).

For the statistical analysis, we used a 2×2 mixed design ANOVA, with age group as between subject variable (with levels "young" and "old") and hand(edness) as within subject variable (with levels "dominant" and "non-dominant"). The significance level p for the statistical analysis was set at 0.05. Post-hoc analysis was performed with the Bonferroni test. The analysis was done with SPSS version 20.

4.2.1 Results

Table 4.2 and Table 4.3 show a summary of the descriptive statistics for the different performance measures in the motor tests and the results of the analysis. In the following sections, we report the results separately for the different test and for each performance measure.

Circle-A

Error

The analysis of Error for the Circle-A motor test revealed a main effect for hand ($F(1, 4135) = 106.97$, $p < 0.001$), with both young adults and old adults making larger errors with their non-dominant hand, a main effect for age group ($F(1, 4135) = 112.13$, $p < 0.001$), with old adults making larger errors than young adults, and no interaction effect ($p = 0.088$).

Speed

Speed analysis revealed a main effect for hand ($F(1, 4135) = 100.92$, $p < 0.001$), with users being faster with their dominant hand in both age groups, and a main effect for age group ($F(1, 4135) = 12.10$, $p < 0.001$), with old adults being faster than young adults. We also found an interaction effect between hand and age group ($F(1, 4135) =$

Table 4.2: Descriptive statistics (mean and standard deviation (sd)) of the performance measures for the two age groups (young adults (18-30) and old adults (50-75))

| Measure | AgeGroup | Dominant Hand | Non-Dominant hand |
|---------------------|----------|-------------------|-------------------|
| | | (mean \pm sd) | (mean \pm sd) |
| Circle-A | | | |
| Error | Young | 0.017 \pm 0.022 | 0.022 \pm 0.018 |
| | Old | 0.027 \pm 0.044 | 0.034 \pm 0.045 |
| Speed | Young | 3.79 \pm 2.56 | 3.49 \pm 2.04 |
| | Old | 4.27 \pm 2.39 | 3.8 \pm 1.99 |
| Movement Smoothness | Young | -14.44 \pm 2.43 | -14.64 \pm 2.32 |
| | Old | -14.36 \pm 2.58 | -14.82 \pm 2.50 |
| Square | | | |
| Error | Young | 0.012 \pm 0.01 | 0.009 \pm 0.009 |
| | Old | 0.013 \pm 0.016 | 0.009 \pm 0.011 |
| Speed | Young | 4.41 \pm 1.99 | 3.89 \pm 1.76 |
| | Old | 4.31 \pm 1.92 | 3.92 \pm 1.68 |
| Movement Smoothness | Young | -14.10 \pm 1.86 | -14.59 \pm 1.9 |
| | Old | -14.64 \pm 2.23 | -14.99 \pm 2.23 |
| Circle-S | | | |
| Error | Young | 0.18 \pm 3.31 | 0.11 \pm 0.073 |
| | Old | 0.064 \pm 0.039 | 0.09 \pm 0.06 |
| Speed | Young | 21.52 \pm 9.17 | 16.3 \pm 6.89 |
| | Old | 17.79 \pm 8.4 | 14.51 \pm 6.08 |
| Movement Smoothness | Young | -17.84 \pm 0.92 | -17.77 \pm 0.89 |
| | Old | -18.02 \pm 1.65 | -18.01 \pm 1.6 |
| Path | | | |
| Error | Young | 0.11 \pm 0.09 | 0.12 \pm 0.09 |
| | Old | 0.16 \pm 0.15 | 0.15 \pm 0.14 |
| Speed | Young | 9.62 \pm 3.75 | 9.09 \pm 3.39 |
| | Old | 7.97 \pm 2.79 | 7.44 \pm 2.40 |
| Movement Smoothness | Young | -11.47 \pm 1.12 | -11.61 \pm 1.08 |
| | Old | -12.35 \pm 1.84 | -12.58 \pm 1.84 |
| Tapping- 2 | | | |
| Reaction Time | Young | 0.26 \pm 0.11 | 0.27 \pm 0.10 |
| | Old | 0.27 \pm 0.11 | 0.28 \pm 0.10 |
| Error | Young | 0.88 \pm 0.96 | 0.91 \pm 0.96 |
| | Old | 1.36 \pm 1.23 | 1.38 \pm 1.23 |
| Tapping-4 | | | |
| Reaction Time | Young | 0.46 \pm 0.06 | 0.47 \pm 0.07 |
| | Old | 0.46 \pm 0.06 | 0.47 \pm 0.06 |
| Error | Young | 0.96 \pm 1.04 | 0.98 \pm 1.02 |
| | Old | 1.22 \pm 1.1 | 1.23 \pm 1.08 |

Table 4.3: Main effects and interaction effects for age group and hand for the different performance measures. (Significant differences ($p < 0.05$) in bold)

| Measure | | | | | | |
|---------------------|-----------|--------------|--------|--------------|-----------------|--------------|
| | Age Group | | Hand | | Age Group *Hand | |
| | F | P | F | P | F | P |
| Circle-A | | | | | | |
| Error | 112.13 | 0.00 | 106.97 | 0.00 | 2.92 | 0.08 |
| Speed | 12.10 | 0.001 | 100.92 | 0.00 | 4.36 | 0.04 |
| Movement Smoothness | 0.17 | 0.68 | 104.76 | 0.00 | 16.27 | 0.00 |
| Square | | | | | | |
| Error | 4.0 | 0.04 | 106.47 | 0.00 | 5.77 | 0.02 |
| Speed | 0.09 | 0.76 | 245.01 | 0.00 | 5.09 | 0.02 |
| Movement Smoothness | 23.19 | 0.00 | 269.45 | 0.00 | 6.33 | 0.01 |
| Circle-S | | | | | | |
| Error | 0.41 | 0.52 | 0.03 | 0.86 | 0.22 | 0.64 |
| Speed | 28.52 | 0.00 | 386.43 | 0.00 | 20.18 | 0.00 |
| Movement Smoothness | 8.20 | 0.004 | 8.46 | 0.004 | 3.80 | 0.05 |
| Path | | | | | | |
| Error | 76.67 | 0.00 | 0.88 | 0.35 | 3.35 | 0.07 |
| Speed | 95.06 | 0.00 | 69.37 | 0.00 | 0.02 | 0.89 |
| Movement Smoothness | 239.82 | 0.00 | 94.23 | 0.00 | 7.01 | 0.008 |
| Tapping- 2 | | | | | | |
| Reaction Time | 2.15 | 0.14 | 23.17 | 0.00 | 0.86 | 0.35 |
| Error | 76.64 | 0.00 | 109.84 | 0.00 | 3.18 | 0.07 |
| Tapping-4 | | | | | | |
| Reaction Time | 0.059 | 0.80 | 162.19 | 0.00 | 0.28 | 0.59 |
| Error | 19.62 | 0.00 | 37.45 | 0.00 | 0.68 | 0.41 |

4.36 , $p < 0.05$). The post-hoc analysis revealed a statistically significant difference between young adults and old adults for both dominant ($p < 0.001$) and non-dominant hand ($p < 0.005$), confirming the main effect of age group. We also found a statistically significant difference between dominant and non-dominant hand for both old adults and young adults ($p < 0.001$), with speed being higher for tests made with the dominant hand, confirming the main effect for hand.

Movement smoothness

The analysis of movement smoothness revealed a main effect for hand ($F(1, 4135) = 104.76$, $p < 0.001$), with movement being less smooth with the non-dominant hand for both age groups, and no main effect for age group ($p = 0.68$). We also found an interaction effect between hand and age-group ($F(1, 4135) = 16.26$, $p < 0.001$). The post-hoc analysis revealed no statistically significant difference between young adults and old adults for dominant ($p = 0.51$) and for non-dominant hand ($p = 0.13$). However, there was a statistically significant difference between dominant and non-dominant hand for both age groups ($p < 0.001$), with movement smoothness being higher for tests made with the dominant hand, confirming the main effect for hand.

Square

Error

Error analysis for the Square motor test revealed a main effect for hand ($F(1, 3892) = 106.47$, $p < 0.001$), with users making larger errors with their dominant hand, and a main effect for age group ($F(1, 3892) = 4.0$, $p < 0.05$), with old adults making larger errors than young adults. We also found an interaction effect between hand and age group ($F(1, 3892) = 5.77$, $p < 0.05$). The post-hoc analysis revealed a statistically significant difference between young adults and old adults for the dominant

hand ($p < 0.05$), and no difference for the non-dominant hand ($p = 0.57$). Moreover, there was a statistically significant difference between dominant and non-dominant hand for both age groups ($p < 0.001$), confirming the main effect for hand.

Speed

Speed analysis revealed a main effect for hand ($F(1, 3892) = 245.01$, $p < 0.001$), with speed being lower with the non-dominant hand, and no main effect for age group ($p = 0.76$). We also found an interaction effect between hand and age group ($F(1, 3892) = 5.09$, $p < 0.05$). The post-hoc analysis revealed no statistically significant difference between young adults and old adults for both dominant ($p = 0.36$) and non-dominant hand ($p = 0.69$). However, there was a statistically significant difference between dominant and non-dominant hand for both age groups ($p < 0.001$), confirming the main effect for hand.

Movement smoothness

Movement smoothness analysis revealed a main effect for hand ($F(1, 3892) = 269.44$, $p < 0.001$), with users making smoother movements with their dominant hand, a main effect for age group ($F(1, 3892) = 23.19$, $p < 0.001$), with smoothness for young adults being higher than for old adults. We also found an interaction effect between hand and age group ($F(1, 3892) = 6.33$, $p < 0.05$). The post-hoc analysis revealed a statistically significant difference between young adults and old adults for both the dominant hand ($p < 0.001$) and the non-dominant hand ($p < 0.001$), with smoothness being higher for the young group, confirming the main effect for age group. We also found a statistically significant difference between dominant and non-dominant hand for both age groups ($p < 0.001$), with movement with the dominant hand being smoother than movement with the non-dominant hand, confirming the main effect for hand.

Circle-S

Error

The analysis of Error for the Circle-S motor test revealed no main effect for hand ($p < 0.86$), no main effect for age group ($p < 0.52$), and no interaction effect ($p = 0.639$).

Speed

Speed analysis revealed a main effect for hand ($F(1, 1340) = 386.43$, $p < 0.001$), with speed being lower with the non-dominant hand, and a main effect for age group ($F(1, 1340) = 28.52$, $p < 0.001$), with speed for young adults being higher than for old adults. We also found an interaction effect between hand and age group ($F(1, 1340) = 20.18$, $p < 0.001$). The post-hoc analysis revealed a statistically significant difference between young adults and old adults for both the dominant hand ($p < 0.001$) and the non-dominant hand ($p < 0.001$), with speed being higher for young adults, confirming the main effect for age. We also found a statistically significant difference between dominant and non-dominant hand for both age groups ($p < 0.001$), with higher speed for the dominant hand than for the non-dominant hand, confirming the main effect for hand.

Movement smoothness

Movement smoothness analysis revealed a main effect for hand ($F(1, 1340) = 8.46$, $p < 0.005$), with users making smoother movements with their dominant hand, a main effect for age group ($F(1, 1340) = 8.2$, $p < 0.005$), with smoothness for young adults being higher than for old adults, and no interaction effect ($p = 0.051$).

Path

Error

The analysis of Error for the Path motor test revealed no main effect for hand ($p = 0.355$), a main effect for age group ($F(1, 3835) = 76.67$, $p < 0.001$), with old adults making larger errors, and no interaction effect ($p = 0.07$).

Speed

Speed analysis revealed a main effect for hand ($F(1, 3835) = 69.37$, $p < 0.001$), with speed being lower with the non-dominant hand, a main effect for age group ($F(1, 3835) = 95.06$, $p < 0.001$), with speed for young adults being higher than for old adults, and no interaction effect ($p = 0.89$).

Movement smoothness

Movement smoothness analysis revealed a main effect for hand ($F(1, 3835) = 94.23$, $p < 0.001$), with users making smoother movements with their dominant hand, and a main effect for age group ($F(1, 3835) = 239.82$, $p < 0.001$), with smoothness for young adults being higher than for old adults. We also found an interaction effect between hand and age-group ($F(1, 3835) = 7.01$, $p < 0.05$). The post-hoc analysis revealed a statistically significant difference between young adults and old adults for both the dominant hand ($p < 0.001$) and the non-dominant hand ($p < 0.001$), with smoothness being higher for the young group, confirming the main effect for age group. We also found a statistically significant difference between dominant and non-dominant hand for both age groups ($p < 0.001$), with movement with the dominant hand being smoother than movement with the non-dominant hand, confirming the main effect for hand.

Tapping-2

Reaction time

Analysis of the Reaction Time for the Tapping-2 motor test revealed a main effect for hand ($F(1, 3346) = 23.70$, $p < 0.001$), with reaction time being higher for the non-dominant hand, no main effect for age group ($p = 0.14$), and no interaction effect ($p = 0.35$).

Error

Analysis of Error revealed a main effect for hand ($F(1, 3346) = 109.84$, $p < 0.001$), with error being larger with the non-dominant hand, a main effect for age group ($F(1, 3346) = 76.64$, $p < 0.001$), with error being larger for old adults, and no interaction effect ($p = 0.074$).

Tapping-4

Reaction time

Analysis of the Reaction Time for the Tapping-4 motor test revealed a main effect for hand ($F(1, 2929) = 162.19$, $p < 0.001$), with reaction time being higher for the non-dominant hand, no main effect for age group ($p = 0.80$), and no interaction effect ($p = 0.59$).

Error

Analysis of Error revealed a main effect for hand ($F(1, 2929) = 37.45$, $p < 0.001$), with users making larger errors with their non-dominant hand, a main effect for age group ($F(1, 2929) = 19.62$, $p < 0.001$), with error being larger for old adults, and no interaction effect ($p = 0.41$).

4.2.2 Discussion

The goal of this analysis was to investigate the patterns of human motor performance in the MotorBrain neuro-motor dataset, which was collected in an unsupervised way from (supposedly) healthy subjects, and compare them with the related literature and the previous controlled study.

As a whole, the results of the analysis seem to be consistent with the literature on human motor performance of the upper limbs and in agreement with the findings of the previous controlled study. In all motor tests, old adults were slower, made larger errors, and made less smoother movements than young adults. Most of the performance measures revealed the degradation of the central and peripheral nervous and neuromuscular systems that is typical of aging [18, 95, 149]. The age-related decline in motor performance is related to several factors, including increasing muscle weakness, increased musculoskeletal stiffness, decreased conduction velocity of nerve fibers, and decreased muscle elasticity. All of these factors impair muscle activation and movement coordination of finger and wrist, resulting in slower movements and longer reaction times [18, 37, 96, 110, 149].

The other main result of the analysis is that both young adults and old adults performed better with their dominant hand than with their non-dominant hand. Human performance with the non-dominant hand is influenced by many factors and is typically worse than performance with the dominant hand because of the different specializations of the two arms and the need for more corrective movements [176, 86]. A common case for human beings is for the left hand to be the non-dominant hand. Since the left hand is controlled by the right hemisphere, which is also strongly involved in visuo-spatial information processing, there is an additional conflict for resources that results in worse motor performance.

Of the different performance measures we used, movement smoothness seemed to be

particularly effective to reveal performance differences in trail making tests. Smoothness is a characteristic of well-coordinated muscle movements that can be validly measured by log dimensionless jerk [13]. The results we obtained for movement smoothness are consistent with a study by Gulde and Hermsdorfer [65] in which the authors used log dimensionless jerk, among other measures, to compare the smoothness of movements of young and old users. The study revealed that healthy young users perform smoother movements compared to healthy old users.

Speed and error were slightly less effective than movement smoothness in highlighting performance differences and were involved in the only two results that were found to be counterintuitive with respect to the literature on the age-related decline in motor performance. In particular, in the Circle-A motor test, speed was higher for old adults than for young adults. Users were instructed to perform the task as quickly as possible with both hands, which led to more errors with the non-dominant hand. This result might be justified by noting that it might be the effect of a speed-accuracy tradeoff, as described by Fitt's law, which states that there is an inverse relationship between accuracy and speed of human motor performance [57, 169]. For example, in a study by Barral et al. [16], the analysis of the movement patterns of 5-year-old boys and girls revealed that speed was higher with the non-dominant hand but accuracy was lower, which was likely due to the speed-accuracy trade off. Since accurate movements require more concentration and more time, old adults achieved faster movements in the test at the expense of higher error, which was indeed one of the results of the analysis. Unfortunately, no supportive evidence for this effect could be obtained by the other trail making test in the accuracy category, Square, for which speed analysis did not reveal any effect for age group.

The other counterintuitive result we obtained concerned the analysis of error in the Square motor test. Here, we found that both young and old adults made larger errors with their dominant hand than with their non-dominant hand. In the literature [21, 141],

it has been reported that the non-dominant hand, as a less skilled hand, shows better performance in writing large letters or drawing shapes. In the case of a square, the users have to rapidly change the direction of movement at the corners, which may cause some hesitation. The smaller errors with the non-dominant hand could be due to the fact that users made more adjustments when performing the square task, due to the complexity of the trajectory to draw.

Another explanation for both counterintuitive results is to call into question the design of the application itself: Circle-A and Square are the first two motor tests that users interact with when using the app. Unfamiliarity might then have played a role and negatively affected performance. A misunderstanding of the instructions for the two tests, especially by old adults, might also be a possible cause.

Finally, reaction time did not appear to be effective in revealing expected age-related performance differences in the two tapping tests. This is partially in agreement with the results of the controlled study, where a similar result was found for Tapping-4 (but a difference in reaction time between young adults and old adults was found for Tapping-2). In this case, it is possible that the specific designs of the two tests are not ideal to reveal age-related performance differences and that different designs should be investigated.

Overall, the analysis provided supporting evidence for the meaningfulness of the data collected in an unsupervised non-clinical way through the MotorBrain app and for the appropriateness of the included motor tests and performance measures.

4.3 Fine-grained age-based data analysis

After the analysis on young adults vs old adults provided evidence that the MotorBrain dataset contains representative data in terms of normative human motor performance, we focused on a more fine-grained analysis based on age. We split the dataset in 5

groups based on the following age ranges, which were chosen in collaboration with the neurologists who worked in the MotorBrain project: 18–27 years old (Group-A), 28–37 years old (Group-B), 38–47 years old (Group-C), 48–57 years old (Group-D), and > 58 years old (Group-E). The goal was to investigate if such a finer granularity could lead to the identification of more detailed information on the degradation of human motor performance with aging.

As in the previous analysis, the number of subjects involved in this analysis in the 5 age groups differs by motor test and is reported in Table 4.4.

Table 4.4: Number of users by the group for the different motor tests.

| MotorBrain Tests | Age Groups | | | | |
|-------------------------|-------------------|----------------|----------------|----------------|----------------|
| | Group-A | Group-B | Group-C | Group-D | Group-E |
| Circle-A | 3233 | 1215 | 693 | 362 | 137 |
| Square | 3038 | 1144 | 643 | 346 | 136 |
| Circle-S | 908 | 497 | 346 | 215 | 75 |
| Path | 2967 | 1157 | 660 | 358 | 137 |
| Tapping-2 | 2584 | 1007 | 602 | 327 | 126 |
| Tapping-4 | 2232 | 894 | 542 | 304 | 122 |

As before, we performed an additional filtering step on the data, keeping only the first run for each motor test, i.e., the data about the first 3 repetitions of a test with each hand.

The analysis concerned the effects of handedness and age, and focused on the same performance measures used previously: error, speed and movement smoothness for trail making tests (Circle-A, Square, Path, Circle-S), and reaction time and error for finger tapping tests (Tapping-2 and Tapping-4).

For the statistical analysis, we used a 2×5 mixed design ANOVA, with age group as between subject variable (with levels "Group-A", "Group-B", "Group-C", "Group-D", "Group-E") and hand(edness) as within subject variable (with levels "dominant" and

"non-dominant"). The significance level p for the statistical analysis was set at 0.05. Post-hoc analysis was performed with the Bonferroni test. The analysis was done with SPSS version 20.

4.3.1 Results

Figures 4.1 to 4.6 show charts of the different performance measures for each age group in the individual motor tests. Table 4.5 shows the results of the analysis in terms of main and interaction effects while Table 4.6 and Table 4.7 show the post-hoc pairwise comparison between groups and between dominant/non-dominant hand after significant interaction effects. In the following sections, we report the results separately for the different motor tests and for each performance measure.

Circle-A

Error

The analysis of Error for the Circle-A motor test revealed a main effect for hand ($F(1, 5635) = 107.49$, $p < 0.001$), with users making larger errors with their non-dominant hand, a main effect for age group ($F(4, 5635) = 39.18$, $p < 0.001$), and no interaction effect ($p = 0.25$). The post-hoc analysis on age group revealed a statistically significant difference between Group-A and Group-C ($p < 0.05$), Group-A and Group-D ($p < 0.05$), Group-A and Group-E ($p < 0.05$), Group-B and Group-D ($p < 0.001$), Group-B and Group-E ($p < 0.001$), Group-C and Group-D ($p < 0.001$), Group-C and Group-E ($p < 0.001$), and Group-D and Group-E ($p < 0.05$), with error being larger for the last group in each pair.

Table 4.5: Main effects and interaction effects for age group and hand for the different performance measures. (Significant differences ($p < 0.05$) in bold)

| Measure | | | | | | |
|---------------------|-----------|--------------|--------|--------------|-----------------|--------------|
| | Age Group | | Hand | | Age Group *Hand | |
| | F | P | F | P | F | P |
| Circle-A | | | | | | |
| Error | 39.18 | 0.00 | 107.49 | 0.00 | 1.35 | 0.25 |
| Speed | 11.38 | 0.00 | 154.28 | 0.00 | 2.92 | 0.02 |
| Movement Smoothness | 13.94 | 0.00 | 177.18 | 0.00 | 7.41 | 0.00 |
| Square | | | | | | |
| Error | 2.03 | 0.09 | 164.28 | 0.00 | 3.11 | 0.01 |
| Speed | 2.82 | 0.02 | 400.83 | 0.00 | 1.53 | 0.19 |
| Movement Smoothness | 13.70 | 0.00 | 438.24 | 0.00 | 2.06 | 0.08 |
| Circle-S | | | | | | |
| Error | 0.70 | 0.59 | 0.03 | 0.87 | 0.79 | 0.53 |
| Speed | 8.41 | 0.00 | 467.58 | 0.00 | 8.17 | 0.00 |
| Movement Smoothness | 2.99 | 0.02 | 8.97 | 0.003 | 1.78 | 0.13 |
| Path | | | | | | |
| Error | 20.33 | 0.00 | 0.70 | 0.40 | 0.77 | 0.55 |
| Speed | 31.02 | 0.00 | 109.38 | 0.00 | 4.16 | 0.002 |
| Movement Smoothness | 72.29 | 0.00 | 142.57 | 0.00 | 4.47 | 0.001 |
| Tapping- 2 | | | | | | |
| Reaction Time | 0.39 | 0.82 | 13.61 | 0.00 | 0.46 | 0.76 |
| Error | 4.26 | 0.002 | 5.67 | 0.02 | 7.14 | 0.00 |
| Tapping-4 | | | | | | |
| Reaction Time | 1.86 | 0.12 | 100.33 | 0.00 | 0.50 | 0.73 |
| Error | 0.67 | 0.61 | 16.79 | 0.00 | 10.45 | 0.00 |

Table 4.6: Pairwise comparisons between groups after significant interaction effects. Significant differences are highlighted in green.

| Task Name | | Speed in Dominant Hand | | | | | | Speed in non-Dominant Hand | | | | | | |
|-----------|-----------------------------|------------------------|---------|---------|---------|---------|---------------------------------|----------------------------|---------|---------|---------|---------|---------|---------|
| A | | Group-A | Group-B | Group-C | Group-D | Group-E | B | | Group-A | Group-B | Group-C | Group-D | Group-E | |
| Circle-A | Group-A | - | 0.000 | 0.001 | 0.010 | 0.160 | Group-A | - | 0.000 | 0.004 | 0.033 | 0.940 | | |
| | Group-B | - | - | 1.000 | 1.000 | 1.000 | Group-B | - | - | 1.000 | 1.000 | 1.000 | | |
| | Group-C | - | - | - | 1.000 | 1.000 | Group-C | - | - | - | 1.000 | 1.000 | | |
| | Group-D | - | - | - | - | 1.000 | Group-D | - | - | - | - | 1.000 | | |
| | Group-E | - | - | - | - | - | Group-E | - | - | - | - | - | | |
| | Smoothness in Dominant Hand | | | | | | Smoothness in non-Dominant Hand | | | | | | | |
| | A | | Group-A | Group-B | Group-C | Group-D | Group-E | B | | Group-A | Group-B | Group-C | Group-D | Group-E |
| | Group-A | - | 0.000 | 0.006 | 0.485 | 1.000 | Group-A | - | 0.000 | 0.001 | 1.000 | 0.131 | | |
| | Group-B | - | - | 1.000 | 0.828 | 0.013 | Group-B | - | - | 1.000 | 0.061 | 0.000 | | |
| | Group-C | - | - | - | 0.657 | 0.011 | Group-C | - | - | - | 0.225 | 0.001 | | |
| Group-D | - | - | - | - | 0.622 | Group-D | - | - | - | - | 0.228 | | | |
| Group-E | - | - | - | - | - | Group-E | - | - | - | - | - | | | |
| Square | Error in Dominant Hand | | | | | | Error in non-Dominant Hand | | | | | | | |
| | A | | Group-A | Group-B | Group-C | Group-D | Group-E | B | | Group-A | Group-B | Group-C | Group-D | Group-E |
| | Group-A | - | 0.933 | 1.000 | 0.589 | 0.057 | Group-A | - | 1.000 | 0.565 | 1.000 | 1.000 | | |
| | Group-B | - | - | 1.000 | 1.000 | 0.423 | Group-B | - | - | 0.643 | 1.000 | 1.000 | | |
| | Group-C | - | - | - | 1.000 | 0.101 | Group-C | - | - | - | 1.000 | 1.000 | | |
| | Group-D | - | - | - | - | 1.000 | Group-D | - | - | - | - | 1.000 | | |
| Group-E | - | - | - | - | - | Group-E | - | - | - | - | - | | | |
| Circle-S | Speed in Dominant Hand | | | | | | Speed in non-Dominant Hand | | | | | | | |
| | A | | Group-A | Group-B | Group-C | Group-D | Group-E | B | | Group-A | Group-B | Group-C | Group-D | Group-E |
| | Group-A | - | 0.179 | 0.507 | 0.001 | 0.000 | Group-A | - | 1.000 | 1.000 | 1.000 | 0.000 | | |
| | Group-B | - | - | 1.000 | 0.545 | 0.001 | Group-B | - | - | 1.000 | 1.000 | 0.000 | | |
| | Group-C | - | - | - | 0.565 | 0.001 | Group-C | - | - | - | 0.319 | 0.000 | | |
| | Group-D | - | - | - | - | 0.109 | Group-D | - | - | - | - | 0.014 | | |
| Group-E | - | - | - | - | - | Group-E | - | - | - | - | - | | | |
| Path | Speed in Dominant Hand | | | | | | Speed in non-Dominant Hand | | | | | | | |
| | A | | Group-A | Group-B | Group-C | Group-D | Group-E | B | | Group-A | Group-B | Group-C | Group-D | Group-E |
| | Group-A | - | 0.331 | 0.010 | 0.000 | 0.000 | Group-A | - | 0.000 | 0.001 | 0.000 | 0.000 | | |
| | Group-B | - | - | 1.000 | 0.000 | 0.000 | Group-B | - | - | 1.000 | 0.000 | 0.000 | | |
| | Group-C | - | - | - | 0.003 | 0.000 | Group-C | - | - | - | 0.000 | 0.000 | | |
| | Group-D | - | - | - | - | 0.083 | Group-D | - | - | - | - | 0.258 | | |
| | Group-E | - | - | - | - | - | Group-E | - | - | - | - | - | | |
| | Smoothness in Dominant Hand | | | | | | Smoothness in non-Dominant Hand | | | | | | | |
| | A | | Group-A | Group-B | Group-C | Group-D | Group-E | B | | Group-A | Group-B | Group-C | Group-D | Group-E |
| | Group-A | - | 1.000 | 0.127 | 0.000 | 0.000 | Group-A | - | 0.116 | 0.002 | 0.000 | 0.000 | | |
| Group-B | - | - | 0.738 | 0.000 | 0.000 | Group-B | - | - | 1.000 | 0.000 | 0.000 | | | |
| Group-C | - | - | - | 0.000 | 0.000 | Group-C | - | - | - | 0.000 | 0.000 | | | |
| Group-D | - | - | - | - | 0.000 | Group-D | - | - | - | - | 0.000 | | | |
| Group-E | - | - | - | - | - | Group-E | - | - | - | - | - | | | |
| Tapping-2 | Error in Dominant Hand | | | | | | Error in non-Dominant Hand | | | | | | | |
| | A | | Group-A | Group-B | Group-C | Group-D | Group-E | B | | Group-A | Group-B | Group-C | Group-D | Group-E |
| | Group-A | - | 1.000 | 0.004 | 0.064 | 0.000 | Group-A | - | 1.000 | 1.000 | 1.000 | 1.000 | | |
| | Group-B | - | - | 0.012 | 0.010 | 0.000 | Group-B | - | - | 1.000 | 1.000 | 1.000 | | |
| | Group-C | - | - | - | 1.000 | 0.182 | Group-C | - | - | - | 1.000 | 1.000 | | |
| | Group-D | - | - | - | - | 0.756 | Group-D | - | - | - | - | 1.000 | | |
| Group-E | - | - | - | - | - | Group-E | - | - | - | - | - | | | |
| Tapping-4 | Error in Dominant Hand | | | | | | Error in non-Dominant Hand | | | | | | | |
| | A | | Group-A | Group-B | Group-C | Group-D | Group-E | B | | Group-A | Group-B | Group-C | Group-D | Group-E |
| | Group-A | - | 1.000 | 0.020 | 0.047 | 0.033 | Group-A | - | 1.000 | 1.000 | 1.000 | 1.000 | | |
| | Group-B | - | - | 0.635 | 0.493 | 0.187 | Group-B | - | - | 1.000 | 1.000 | 1.000 | | |
| | Group-C | - | - | - | 1.000 | 1.000 | Group-C | - | - | - | 1.000 | 1.000 | | |
| | Group-D | - | - | - | - | 1.000 | Group-D | - | - | - | - | 1.000 | | |
| Group-E | - | - | - | - | - | Group-E | - | - | - | - | - | | | |

Table 4.7: Pairwise comparisons between dominant and non-dominant hand after significant interaction effects. Significant differences are highlighted in green.

| Task Name | | | |
|---------------------|-----------|---------------|-------------------|
| Circle-A | | | |
| Speed | Age Group | Dominant Hand | non-Dominant Hand |
| | Group-A | 0.000 | 0.000 |
| | Group-B | 0.000 | 0.000 |
| | Group-C | 0.000 | 0.000 |
| | Group-D | 0.000 | 0.000 |
| | Group-E | 0.000 | 0.000 |
| Movement Smoothness | Age Group | Dominant Hand | non-Dominant Hand |
| | Group-A | 0.000 | 0.000 |
| | Group-B | 0.000 | 0.000 |
| | Group-C | 0.000 | 0.000 |
| | Group-D | 0.000 | 0.000 |
| | Group-E | 0.000 | 0.000 |
| Square | | | |
| Error | Age Group | Dominant Hand | non-Dominant Hand |
| | Group-A | 0.000 | 0.000 |
| | Group-B | 0.000 | 0.000 |
| | Group-C | 0.000 | 0.000 |
| | Group-D | 0.000 | 0.000 |
| | Group-E | 0.000 | 0.000 |
| Circle-S | | | |
| Speed | Age Group | Dominant Hand | non-Dominant Hand |
| | Group-A | 0.000 | 0.000 |
| | Group-B | 0.000 | 0.000 |
| | Group-C | 0.000 | 0.000 |
| | Group-D | 0.000 | 0.000 |
| | Group-E | 0.000 | 0.000 |
| Path | | | |
| Speed | Age Group | Dominant Hand | non-Dominant Hand |
| | Group-A | 0.000 | 0.000 |
| | Group-B | 0.000 | 0.000 |
| | Group-C | 0.000 | 0.000 |
| | Group-D | 0.000 | 0.000 |
| | Group-E | 0.057 | 0.057 |
| Movement Smoothness | Age Group | Dominant Hand | non-Dominant Hand |
| | Group-A | 0.000 | 0.000 |
| | Group-B | 0.000 | 0.000 |
| | Group-C | 0.000 | 0.000 |
| | Group-D | 0.000 | 0.000 |
| | Group-E | 0.000 | 0.000 |
| Tapping-2 | | | |
| Error | Age Group | Dominant Hand | non-Dominant Hand |
| | Group-A | 0.000 | 0.000 |
| | Group-B | 0.033 | 0.033 |
| | Group-C | 0.138 | 0.138 |
| | Group-D | 0.030 | 0.030 |
| Tapping-4 | | | |
| Error | Age Group | Dominant Hand | non-Dominant Hand |
| | Group-A | 0.000 | 0.000 |
| | Group-B | 0.439 | 0.439 |
| | Group-C | 0.001 | 0.001 |
| | Group-D | 0.014 | 0.014 |
| | Group-E | 0.000 | 0.000 |

Speed

Speed analysis revealed a main effect for hand ($F(1,5635) = 154.28$, $p < 0.001$), with users being faster with their dominant hand, and a main effect for age group

($F(4, 5635) = 11.39$, $p < 0.001$). We also found an interaction effect between hand and age-groups ($F(4, 5635) = 2.92$, $p < 0.05$). The post-hoc analysis revealed a statistically significant difference between Group-A and Group-B ($p < 0.001$), Group-A and Group-C ($p < 0.005$), and Group-A and Group-D ($p < 0.05$) for both the dominant and non-dominant hand. Moreover, there was a statistically significant difference between dominant and non-dominant hand for each group ($p < 0.001$), confirming the main effect for hand.

Movement smoothness

The analysis of movement smoothness revealed a main effect for hand ($F(1, 5635) = 177.18$, $p < 0.001$), with movement being less smooth with the non-dominant hand, and a main effect for age group ($F(4, 5635) = 13.94$, $p < 0.001$). We also found an interaction effect between hand and age-group ($F(4, 5635) = 7.41$, $p < 0.001$). The post-hoc analysis revealed a statistically significant difference between Group-A and Group-B ($p < 0.001$), Group-A and Group-C ($p < 0.001$), Group-B and Group-E ($p < 0.05$), and Group-C and Group-E ($p < 0.05$) for both the dominant and non-dominant hand. We also found a statistically significant difference between dominant and non-dominant hand for each group ($p < 0.001$), confirming the main effect for hand.

Square

Error

Error analysis for the Square motor test revealed a main effect for hand ($F(1, 5302) = 164.28$, $p < 0.001$), with users making larger errors with their dominant hand, and no main effect for age group ($p < 0.088$). We also found an interaction effect between hand and age-group ($F(4, 5302) = 3.11$, $p < 0.05$). The post-hoc analysis revealed a statistically significant difference between dominant and non-dominant hand for each



Figure 4.1: Mean values of performance measures in Circle-A task by age group

group ($p < 0.001$), confirming the main effect for hand.

Speed

Speed analysis revealed a main effect for hand ($F(1, 5302) = 400.83$, $p < 0.001$), with speed being lower with the non-dominant hand, a main effect for age group ($F(4, 5302) = 2.82$, $p < 0.05$), and no interaction effect between hand and age group ($p < 0.190$). The post-hoc analysis on age group revealed a statistically significant difference between Group-A and Group-B ($p < 0.05$).

Movement smoothness

Movement smoothness analysis revealed a main effect for hand ($F(1, 5302) = 438.24$, $p < 0.001$), with users making smoother movements with their dominant hand, a main effect for age group ($F(4, 5302) = 13.7$, $p < 0.001$), and no interaction effect ($p < 0.083$). The post-hoc analysis on the age group revealed a statistically significant difference be-

tween Group-A and Group-B ($p < 0.005$), Group-A and Group-E ($p < 0.001$), Group-B and Group-D ($p < 0.001$), Group-B and Group-E ($p < 0.001$), Group-C and Group-E ($p < 0.001$), and Group-D and Group-E ($p < 0.05$).

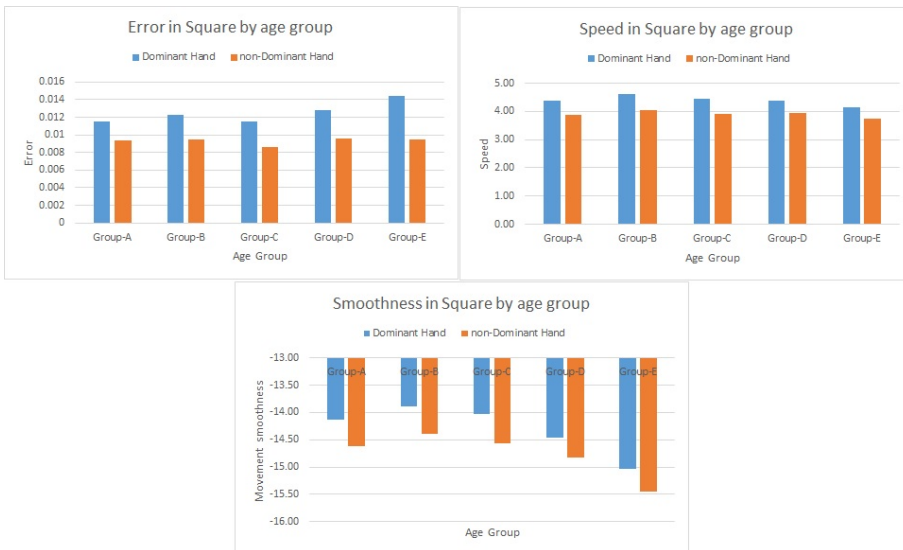


Figure 4.2: Mean values of performance measures in the Square test by age group

Circle-S

Error

The analysis of Error for the Circle-S motor test revealed no main effect for hand ($p < 0.87$), no main effect for age group ($p < 0.59$), and no interaction effect ($p = 0.53$).

Speed

Speed analysis revealed a main effect for hand ($F(1, 2036) = 467.58$, $p < 0.001$), with speed being lower with the non-dominant hand, and a main effect for age group ($F(4, 2036) = 8.4$, $p < 0.001$). We also found an interaction effect between hand and age-group ($F(4, 2036) = 8.17$, $p < 0.001$). The post-hoc analysis revealed a

statistically significant difference in speed between Group-A and Group-D ($p < 0.001$), Group-A and Group-E ($p < 0.001$), Group-B and Group-E ($p < 0.001$), and Group-C and Group-E ($p < 0.001$) for the dominant hand, and between Group-A and Group-E ($p < 0.001$), Group-B and Group-E ($p < 0.001$), Group-C and Group-E ($p < 0.001$), and Group-D and Group-E ($p < 0.05$) for the non-dominant hand. In addition, we found a statistically significant difference between dominant and non-dominant hand for each group ($p < 0.001$), confirming the main effect for hand.

Movement smoothness

Movement smoothness analysis revealed a main effect for hand ($F(1, 2036) = 8.97, p < 0.005$), with users making smoother movements with the dominant hand, a main effect for age group ($F(4, 2036) = 2.99, p < 0.05$), and no interaction effect ($p = 0.130$). However, the post-hoc analysis on age group did not reveal any statistically significant difference between groups.



Figure 4.3: Mean values of performance measures in the Circle-S test by age group

Path

Error

The analysis of Error for the Path motor test revealed no main effect for hand ($p = 0.402$), a main effect for age group ($F(4, 5274) = 20.33$, $p < 0.001$), and no interaction effect ($p = 0.548$). The post-hoc analysis on age group revealed a statistically significant difference between Group-A and Group-D ($p < 0.001$), Group-A and Group-E ($p < 0.001$), Group-B and Group-D ($p < 0.001$), Group-B and Group-E ($p < 0.001$), Group-C and Group-E ($p < 0.001$), and Group-D and Group-E ($p < 0.005$)

Speed

Speed analysis revealed a main effect for hand ($F(1, 5274) = 109.38$, $p < 0.001$), with speed being lower with the non-dominant hand, and a main effect for age group ($F(4, 5274) = 31.02$, $p < 0.001$). We also found an interaction effect between hand and age-group ($F(4, 5274) = 4.16$, $p < 0.005$). The post-hoc analysis revealed a statistically significant difference in speed between Group-A and Group-C ($p < 0.05$), Group-A and Group-D ($p < 0.001$), Group-A and Group-E ($p < 0.001$), Group-B and Group-D ($p < 0.001$), Group-B and Group-E ($p < 0.001$), Group-C and Group-D ($p < 0.005$), and Group-C and Group-E ($p < 0.001$) for the dominant hand, and between Group-A and Group-B ($p < 0.001$), Group-A and Group-C ($p < 0.005$), Group-A and Group-D ($p < 0.001$), Group-A and Group-E ($p < 0.001$), Group-B and Group-D ($p < 0.001$), Group-B and Group-E ($p < 0.001$), Group-C and Group-D ($p < 0.001$), and Group-C and Group-E ($p < 0.001$) for the non-dominant hand. Moreover, there was a statistically significant difference between dominant and non-dominant hand for each group ($p < 0.001$), except for Group-E.

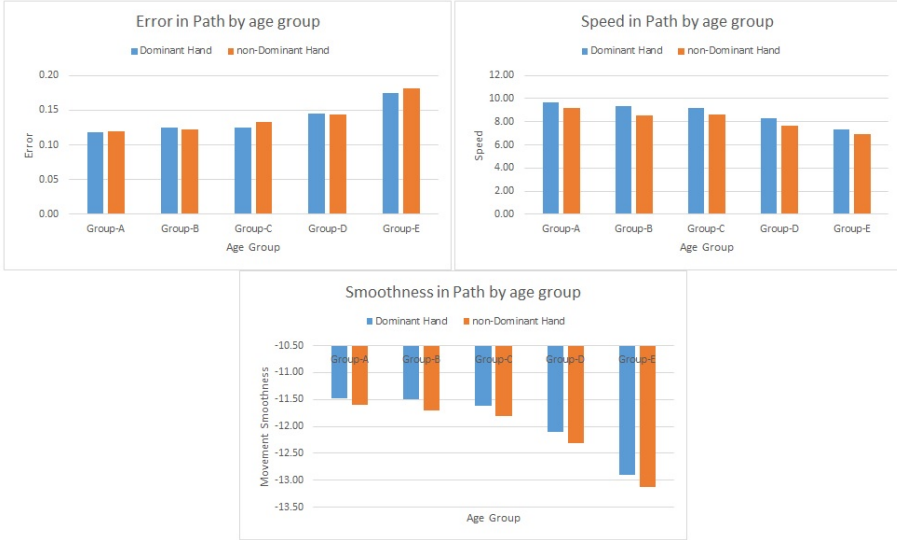


Figure 4.4: Mean values of performance measures in the Path test by age group

Movement smoothness

Movement smoothness analysis revealed a main effect for hand ($F(1, 5274) = 142.57, p < 0.001$), with users making smoother movements with their dominant hand, and a main effect for age group ($F(4, 5274) = 72.29, p < 0.001$). We also found an interaction effect between hand and age-group ($F(4, 5274) = 4.47, p < 0.005$). The post-hoc analysis revealed a statistically significant difference in movement smoothness between Group-A and Group-D ($p < 0.001$), Group-A and Group-E ($p < 0.001$), Group-B and Group-D ($p < 0.001$), Group-B and Group-E ($p < 0.001$), Group-C and Group-D ($p < 0.001$), Group-C and Group-E ($p < 0.001$), and Group-D and Group-E ($p < 0.001$), for the dominant hand, and between Group-A and Group-C ($p < 0.005$), Group-A and Group-D ($p < 0.001$), Group-A and Group-E ($p < 0.001$), Group-B and Group-D ($p < 0.001$), Group-B and Group-E ($p < 0.001$), Group-C and Group-D ($p < 0.001$), Group-C and Group-E ($p < 0.001$), and Group-D and Group-E ($p < 0.001$) for the non-dominant hand. We also found a statistically significant difference between dominant and non-

dominant hand for each group ($p < 0.001$), confirming the main effect for hand.

Tapping-2

Reaction time

Analysis of the reaction time for the Tapping-2 motor test revealed a main effect for hand ($F(1, 4641) = 13.61$, $p < 0.001$), no main effect for age group ($p = 0.818$), and no interaction effect ($p = 0.764$).

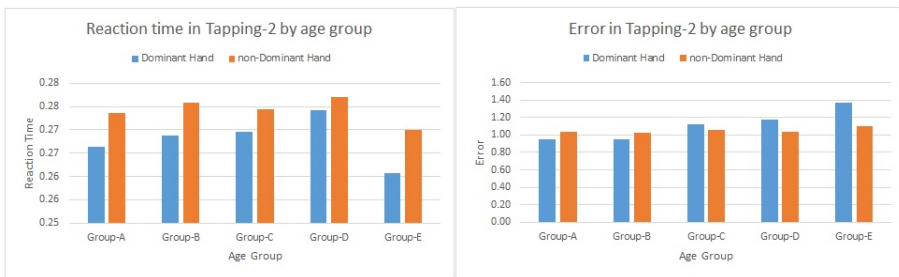


Figure 4.5: Mean values of performance measures in the Tapping-2 test by age group

Error

Analysis of the error revealed a main effect for hand ($F(1, 4641) = 5.67$, $p < 0.05$) and a main effect for age group ($F(4, 4641) = 4.26$, $p < 0.005$). We also found an interaction effect between hand and age-group ($F(4, 4641) = 7.15$, $p < 0.001$). The post-hoc analysis revealed a statistically significant difference in error between Group-A and Group-C ($p < 0.005$), Group-A and Group-D ($p < 0.005$), Group-A and Group-E ($p < 0.001$), Group-B and Group-C ($p < 0.05$), Group-B and Group-D ($p < 0.01$), and Group-B and Group-E ($p < 0.001$) for the dominant hand. In addition, we found a statistically significant difference between dominant and non-dominant hand for Group-A ($p < 0.001$) and for Group-B ($p < 0.05$), both with error being higher with the non-dominant hand, and for Group-D ($p < 0.05$) and Group-E ($p < 0.005$), both with

error being higher with the dominant hand.

Tapping-4

Reaction time

Analysis of the reaction time for the Tapping-4 motor test revealed a main effect for hand ($F(1, 4089) = 100.33$, $p < 0.001$), with reaction time being higher for the non-dominant hand, no main effect for age group ($p = 0.115$), and no interaction effect ($p = 0.731$).

Error

Analysis of the error revealed a main effect for hand ($F(1, 4089) = 16.79$, $p < 0.001$) and no main effect for age group ($p < 0.612$). We also found an interaction effect between hand and age-group ($F(4, 4089) = 10.45$, $p < 0.001$). The post-hoc analysis revealed a statistically significant difference in error between Group-A and Group-C ($p < 0.05$), Group-A and Group-D ($p < 0.05$), and Group-A and Group-E ($p < 0.05$) for the dominant hand. In addition, there was a statistically significant difference between dominant and non-dominant hand for Group-A ($p < 0.001$), with error being higher with the non-dominant hand, and for Group-C ($p < 0.001$), Group-D ($p < 0.05$) and Group-E ($p < 0.001$), all with error being higher with the dominant hand.

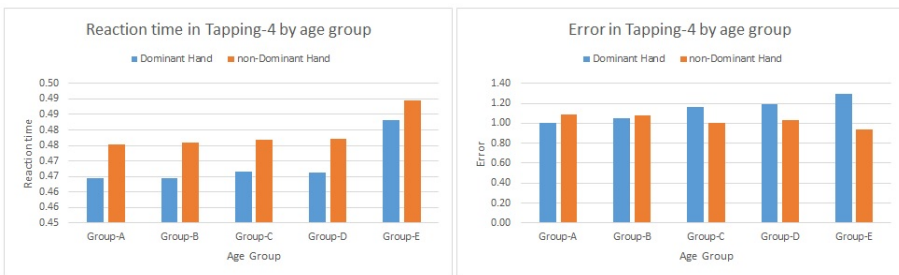


Figure 4.6: Mean values of performance measures in the Tapping-4 test by age group

4.3.2 Discussion

Overall, the results of this fine-grained analysis are largely consistent with the results of the previous analysis and with the literature on human motor performance of the upper limbs. In all motor tests, user performance with the dominant hand was better than performance with the non-dominant hand. This is in agreement with the asymmetry in arm motor performance that humans exhibit, which is associated with specific motor control specializations of each arm, with the dominant arm specialized for predictive control of limb and task dynamics and the non-dominant arm specialized for stabilizing performance [8, 9, 86]. The only major exception to this pattern was found in the Square motor test, where users of all age groups made larger errors with their dominant hand. This is exactly the same result we found in the young adults vs. old adults study. Like before, possible motivations for it can be related to a combination of the speed-accuracy trade-off and the complexity of the movements required by the task, as well as to the design of the test and the application themselves. The dominant vs. non-dominant hand performance pattern was found to be typically consistent across all age groups except for the Tapping-4 test, where error performance of Group-A users was better with the dominant hand but the opposite was true for Group-C, Group-D, and Group-E users. Why performance in this test would become better with the non-dominant hand with increasing age is hard to explain and requires further investigation.

As in the previous analysis, we found strong evidence of the decline of human motor performance with aging, which is due to several factors that negatively affect muscle activation and movement coordination [18, 37, 95, 149, 96, 110, 149]. This is especially clear when comparing the performance of the youngest users (Group-A) with that of the oldest users (Group-E), with the first group being typically faster, more accurate, and making smoother movements than the second group in most tests. The same evidence was often found for Group-B and Group-C with respect to Group-E, showing

that performance degradation was not linear across the age groups. Performance of users in Group-A, Group-B, and Group-C was mostly similar while degradation became evident only for Group-D, typically reaching its maximum for Group-E. In several measures, Group-D showed better performance than Group-E. These results are interesting because they give neurologists the opportunity to study the aging of the population's neuromuscular system [33] across the entire range of ages, while typical results of studies on motor performance and aging are limited to young vs. old groups.

Movement smoothness and speed were slightly more frequently involved than error in highlighting performance differences in trail making tests while error was more effective than reaction time in finger tapping tests. A peculiar effect was revealed in the analysis of movement smoothness in the two motor tests in the accuracy category (Circle-A and Square). The performance curve for smoothness in both tests had a characteristic U-shape, with performance for Group-A being worse than performance for Group-B (and also Group-C for Circle-A). The motivation for this pattern is unclear and requires further investigation. As before, reaction time did not appear to be effective in revealing expected age-related performance differences in the two tapping tests. As reaction time is the defining measure for tapping tests, a different implementation of the tests is probably needed to obtain more useful results.

Overall, the analysis provided further information about age-related motor patterns and additional evidence for the meaningfulness of the data collected by the MotorBrain app and the appropriateness of the included motor tests and performance measures. The dataset could then be used by neurologists as a reference to support the diagnosis of disorders affecting motor control, making it possible to compare the motor performance of new cases with the normative behavior revealed by the existing data. This could also be done with automatic approaches based on machine learning, e.g., by classifying new cases with algorithms trained on the performance patterns of the different age groups.

5

Machine learning models for age-based classification of motor performance

The statistical analysis of the MotorBrain dataset revealed significant differences in motor performance between age groups, with performance differences becoming larger and more frequent the more groups differed in terms of (median) age. This result motivated us to implement machine learning models for the age-based classification of motor performance.

Most studies that employ machine learning in the literature on movement disorders use a supervised approach that is based on training datasets consisting of both healthy subjects and subjects with neurological disorders (see Section 2.3). This makes it possible to directly classify new cases as healthy or not-healthy but is typically based on training data related to subjects with neurological disorders in an advanced stage. Considering that it would be crucial to diagnose a patient in the early stage of a movement

disorder, when symptoms are mild and often ignored, the approach used in current machine learning solutions can result in sub-optimal outcomes. Since the (cleaned) MotorBrain dataset contains only data about (presumably) healthy subjects, our approach was to build machine learning models that classify subjects based on their age group. The idea is that if a new case is misclassified with respect to her known age, such case shows motor patterns that differ from the typical motor patterns of her age group, thus requiring further manual investigation by a neurologist.

In this chapter, we present the machine learning process we followed to build the models, starting with the extraction of appropriate features and continuing with feature selection, training, and evaluation of the results.

5.1 Feature extraction

The set of features we used to support the learning process includes the same performance measures we used in the statistical analysis, augmented with additional features derived from the literature. Table 5.1 shows a complete list of the considered features. In the following, we provide details about the additional features, divided by motor test category.

5.1.1 Trail making features

Four of the motor tests included in MotorBrain (Circle-A, Square, Path, Circle-S) are trail making tests that require users to draw or follow a path on the screen.

In addition to the previously discussed performance measures, we computed 5 new features for each trail-making test: drawing velocity, normalized velocity variability, Shannon entropy, skewness, and kurtosis.

Drawing velocity

Velocity is defined as the rate of change in position while drawing with respect to time and is calculated using the following formula:

$$V = \sum_{i=1}^{N-1} \frac{\sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}}{t_{i+1} - t_i} \quad (5.1)$$

Velocity is also used as input in the computation of the other 4 derived features.

Normalized velocity variability (NVV)

Normalized velocity variability (NVV) describes the subject's drawing velocity variability and has been shown to be effective in identifying abnormal movement patterns. The NVV was introduced by Kotsavasiloglou et al. [90], who used it as a way to capture the balance, or lack thereof, of muscle tone between opposing muscle systems. Indeed, the low-level control of muscle systems occurs on a time scale of the order of milliseconds, whereas the conscious control of movement cannot occur at such a high frequency. The NVV is lower for more regular or smoother movements than for irregular movements. The following formula can be used to calculate this feature:

$$NVV = \frac{1}{T |MV|} \sum_{i=1}^{N-1} |v_{i+1} - v_i| \quad (5.2)$$

Where $MV = \frac{1}{N} \sum_{i=1}^N v_i$, T is the total time, N is the number of data points, and v is the magnitude of the respective velocity data point.

Shannon entropy

Entropy is a measure of a signal's randomness or uncertainty [52]. Shannon entropy reveals hidden complexities of physiological systems in trail making tests. Shannon entropy for the velocity of each repetition is calculated as:

$$ETP_v = - \sum_{i=1}^N P(v_i) \log_2(P(v_i)) \quad (5.3)$$

Where N is the number of data points, v is the velocity of data points, and $P(v)$ is the probability density function calculated by kernel density estimation. To compute Shannon entropy, we used the Sklearn KernelDensity function with a bandwidth of 10 and a Gaussian kernel, with all other parameters using their default values.

Skewness and kurtosis

The use of skewness and kurtosis as features was proposed by Drotár et al [51] to characterize the handwriting patterns of PD patients. Kurtosis measures whether a data distribution is "heavily tailed" or "lightly tailed" compared to a normal distribution, while skewness measures the degree of asymmetry in a data distribution. In the trail-making testing, we calculated skewness and kurtosis of velocity.

5.1.2 Finger tapping features

For the 2 tapping tests, along with the previously defined measures (reaction time and tap precision), we computed 2 new features: tap speed and integrated cognitive assessment.

Tap speed

Tap speed is the number of taps a user makes in the unit of time. In MotorBrain tapping tests, users have only 10 seconds to complete one repetition. Thus, if a user makes N taps during a repetition, then the tap speed is $N/10$.

Table 5.1: Description of features extracted from shapes where {TMT}: trial making tests, {TT}: for tapping tests.

| Shape | Features | Description |
|-------|----------------------------|---|
| {TMT} | mean(error) | The mean error in three repetitions. |
| {TMT} | SD (error) | Standard deviation of error in three repetitions. |
| {TMT} | COV (Error) | The coefficient of variation in error |
| {TMT} | mean (Drawing Velocity) | The mean drawing velocity in three repetitions |
| {TMT} | SD (Drawing Velocity) | Standard deviation of drawing velocity in three repetitions. |
| {TMT} | COV (Drawing Velocity) | The coefficient of variation in Drawing Velocity |
| {TMT} | mean (NVV) | Characterize the variability of drawing velocity in three repetitions. |
| {TMT} | Standard Deviation (NVV) | Standard deviation of variability in drawing velocity in three repetitions |
| {TMT} | Skew (v) | Measures the symmetry of velocity distribution. |
| {TMT} | Kurt (v) | Measure of taildness of velocity distribution. |
| {TMT} | mean (Movement Speed) | The mean of movement speed in three repetitions. |
| {TMT} | SD (Movement Speed) | Standard deviation of movement speed in three repetitions. |
| {TMT} | COV (Movement Speed) | The coefficient of variation in movement speed |
| {TMT} | mean (Movement Smoothness) | The mean movement smoothness in three repetitions |
| {TMT} | SD (Movement Smoothness) | Standard deviation of movement smoothness in three repetitions. |
| {TMT} | Entropy (velocity1) | Measures the disorder or hidden complexities at repetition 1 using the entropy formula. |
| {TMT} | Entropy (velocity2) | Measures the disorder or hidden complexities at repetition 2 using the entropy formula. |
| {TMT} | Entropy (velocity3) | Measures the disorder or hidden complexities at repetition 3 using the entropy formula. |
| {TT} | Mean (rt) | The mean of reaction time in three repetitions. |
| {TT} | SD (rt) | Standard deviation of reaction time in three repetitions. |
| {TT} | Skew (rt) | Measures the symmetry of reaction time |
| {TT} | Kurt (rt) | Measures the taildnes of reaction time distribution |
| {TT} | COV (rt) | The coefficient of variation in reaction time. |
| {TT} | Mean (Tapping Speed) | Mean of (No. of taps in 10 seconds) in three repetitions |
| {TT} | SD (Tapping Speed) | Standard deviation of tapping speed in three repetitions. |
| {TT} | Mean (Error) | The mean error in three repetitions. |
| {TT} | SD (Error) | Standard deviation of error in three repetitions. |
| {TT} | COV (Error) | The coefficient of variation in error |
| {TT} | Mean (ICA) | Mean of Integrated cognitive assessment while tapping. |
| {TT} | SD (ICA) | Standard deviation of ICA. |

Integrated cognitive assessment

Tapping tests involve both the visual and motor cortex, which are affected in the early stages of neurodegenerative diseases. The integrated cognitive assessment (ICA) measures the speed and accuracy of visual and motor processing [85]. ICA is calculated as follows, making use of tap speed and error (one of the previously available measures):

$$ICA = \left(\frac{tapspeed}{100} * \frac{(1 - error)}{100} \right) * 100 \quad (5.4)$$

5.2 Feature selection

In total, the number of features we have is 18 for trail making tests and 12 for tapping tests. In the feature selection stage, we identified the optimal set of features for the machine-learning classification problem, using two steps to remove any redundant or irrelevant features that could cause a classification error. First, we eliminated non-informative features that had a median absolute deviation (MAD) of zero. Second, a subset of discriminatory features was selected using two different feature selection methods: Minimum Redundancy, Maximum Relevance (MRMR) and Recursive Feature Elimination with SVM (RFE-SVM). In both methods, the size of the optimal subset of features was controlled by specifying the number of features (N) in advance.

5.2.1 MRMR

MRMR is a filtering method based on mutual information [98] that aims to identify a small set of features that have the maximum possible predictive power (when used together). It is an efficient approach to select features that are strongly associated with the class labels (response variable). This strong association is captured by the mutual information between the class labels and the feature vector. In the end, the features that are weakly associated with the other features but strongly associated with the class labels are selected. The method is based on the following formula:

$$J_{MRMR}(X_i) = I(Y, X_i) - \frac{1}{|S|} \sum_{X_j \in S} I(X_i, X_j) \quad (5.5)$$

where X_i is a feature that is not in the selected set yet, $J_{MRMR}(X_i)$ is the importance of feature X_i , Y is the response variable (class label), S is the set of currently selected features, $|S|$ is the cardinality of set S , $X_j \in S$ is one of the features in set S . The function $I(Y, X_i)$ is the mutual information between class label Y and feature X_i and measures

the strength of the association between the two variables. Similarly, $I(X_i, X_j)$ is the mutual information between variables X_i and X_j . The second part of the formula thus measures the redundancy of feature X_i . At each step of the MRMR process, the feature with the highest importance score is added to the selected feature set S .

5.2.2 RFE-SVM

The recursive feature elimination (RFE) method is used to find the features that can optimize the performance of a pre-chosen model, e.g. SVM [137]. It explicitly models the correlation between features, resulting in robust classifier performance. It starts with the complete set of features, fitting the model and removing the weakest feature(s) (in terms of scores such as feature importance). The process is repeated until the desired number of features is reached.

Since the desired number of features to keep is not known in advance, cross-validation is used with RFE to score different feature subsets and select the best scoring collection of features. In our case, we tested the effect of using different numbers of features on the predictive power of the two classifiers we employed, Random Forest (RF) and Logistic Regression (LR), finally selecting the features that optimized the predictive power.

5.3 Synthetic data generation

Table 4.1 shows that the number of samples for the young age group is higher than the number of samples for the old age group. Similarly table 4.4 shows that the number of samples in the 5 age groups differ significantly. Classification results can be inaccurate if we train a classifier with imbalanced datasets. To mitigate this problem, we used the Synthetic Minority Oversampling technique (SMOTE), developed by Chawla et al [30], to decrease the imbalance in our datasets. The SMOTE algorithm has been used in a variety of domains, such as breast cancer detection [53], speech recognition [103], and

network intrusion detection systems [34], to solve the problem of imbalanced data.

The technique balances a dataset by synthetically augmenting the minority class (the class with the lowest number of samples) using the k-nearest neighbors approach. More specifically, the following equation is used to synthetically add samples to the minority class:

$$x_{syn} = x_i + (x_{knn} - x_i) * t \quad (5.6)$$

where x_{syn} is a new sample that is generated by the algorithm using the following steps:

1. Identify the k-nearest neighbor x_{knn} of feature vector x_i .
2. Calculate the difference between the feature vector and its k-nearest neighbor.
3. Generate a random number t between 0 and 1 and multiply it by the difference.
4. Generate the new sample by adding the output in step 3 to the feature vector x_i .

5.4 Classification

For classification purposes, we used the Random Forest and Logistic Regression classifiers.

5.4.1 Random Forest

The Random Forest classifier is an ensemble of decision trees [23]. Each decision tree is trained with a different feature set and a different data subset (bootstrapping). Each tree returns a predicted class and RF then returns the class with the majority of votes. This classifier is suitable for our task because it is robust to outliers and its hyperparameter `class_weight` is suitable for unbalance data.

5.4.2 Logistic Regression

Logistic regression is mostly used for binary classification in machine learning. It uses the sigmoid function to classify the data samples. The hypothesis used in logistic regression is as follows

$$h_{\Theta}(x) = g(\Theta_0 + \Theta_1x_1 + \Theta_2x_2 \dots \Theta_nx_n) \quad (5.7)$$

where x_n and Θ_n are the features and parameters used in the sigmoid or logistic function $g(y)$ to predict the class. $g(y)$ is defined as:

$$g(y) = \frac{1}{1 + e^{-y}} \quad (5.8)$$

The output of the sigmoid function is always between 0 and 1. The output of the hypothesis is positive if the output of the sigmoid function is > 0.5 . Again, a loss function is used to see how well the algorithm performs with the weights, represented here as θ . The output of the loss function is a large value if the value predicted by $h_{\theta}(x)$ is far from the true class label y , and its output is small if the predicted output is close to the true class.

The loss function is defined as

$$J(\Theta) = -\frac{1}{m} \cdot \sum_{i=1}^m (y^i \log(h_{\Theta}(x^i)) + (1-y)^i \log(1-h_{\Theta}(x^i))) \quad (5.9)$$

5.5 Experimental setup

To begin, we divided the entire dataset into a training dataset and a test dataset with a stratified split of (60 – 40%) in order to have a good representation of the minority class in the test set. The training dataset was used to build the model, and the test dataset was used to evaluate the model.

In the training data set, each feature vector was normalized by subtracting the mean and dividing by the standard deviation. The features in the test data set were normalized using the mean and standard deviation of the training set. The optimal set of features was selected using the MRMR and RFE feature selection methods. Since the dataset was unbalanced, the SMOTE oversampling technique was used to augment the training data. In the SMOTE method, the 3 nearest neighbors of each sample in the minority class were used to train the model. Samples were generated only for the minority class, while the majority class remained unchanged. Grid search and 5-fold cross-validation were used to train the model. Grid search was used to estimate the number of estimators = 100 – 300 in the Random Forest classifier and the ranges of $C = 0.001 - 100$ in the Logistic Regression classifier. All other parameters were set to default values.

Then we applied each trained classifier to the test set. The whole process was repeated 10 times, each time using a different set of training and test splits by specifying *randomstate = None* in a stratified train test split. In this way, we obtained a better-generalized model for the entire dataset. The 5-fold cross-validation was repeated 10 times, so we obtained a total of 50 trained classifiers. For performance evaluation, we computed the mean accuracy where $accuracy = \frac{\#correctpredictions}{\#testsubjects}$ and mean Area Under the Curve measures over the 50 classifiers to compare their results. For each test, an independent classifier was trained.

The process was repeated two times, the first for the binary classification of young vs. old adults, and the second for the multiclass classification of our 5 age groups.

5.6 Results

5.6.1 Feature selection

We initially extracted a total number of 18 features for each trail making test and 12 features for the tapping tests (as listed in Table 5.1).

For the trail making tests, the MRMR feature selection method selected a total of 12 features out of the possible 18. The RFE-SVM selection method identified 8 features that led to optimal performance of the LR classifier and 12 features for the RF classifier.

For the tapping test, both MRMR and RFE-SVM identified 6 optimal features out of the initial 12 features.

Table 5.2 shows the features selected by MRMR for trail making and tapping tests in each iteration (as we repeat the whole process 10 times with 5-fold cross-validation). No similar list of unique features can be provided for RFE-SVM as the features changed for each iteration of the classification process.

Table 5.2: Top features selected by MRMR: 12 features for trail making tests and 6 for tapping tests ({TMT}: trial making tests, {TT}: for tapping tests).

| Shape | Features | Description |
|-------|----------------------------|---|
| {TMT} | mean (NVV) | Characterize the variability of drawing velocity in three repetitions. |
| {TMT} | Standard Deviation (NVV) | Standard deviation of variability in drawing velocity in three repetitions |
| {TMT} | Skew (v) | Measures the symmetry of velocity distribution. |
| {TMT} | Kurt (v) | Measure of tailness of velocity distribution. |
| {TMT} | mean (Movement Speed) | The mean of movement speed in three repetitions. |
| {TMT} | SD (Movement Speed) | Standard deviation of movement speed in three repetitions. |
| {TMT} | COV (Movement Speed) | The coefficient of variation in movement speed |
| {TMT} | mean (Movement Smoothness) | The mean movement smoothness in three repetitions |
| {TMT} | SD (Movement Smoothness) | Standard deviation of movement smoothness in three repetitions. |
| {TMT} | Entropy (velocity1) | Measures the disorder or hidden complexities at repetition 1 using the entropy formula. |
| {TMT} | Entropy (velocity2) | Measures the disorder or hidden complexities at repetition 2 using the entropy formula. |
| {TMT} | Entropy (velocity3) | Measures the disorder or hidden complexities at repetition 3 using the entropy formula. |
| {TT} | Skew (rt) | Measures the symmetry of reaction time |
| {TT} | Kurt (rt) | Measures the tailness of reaction time distribution |
| {TT} | SD (Tapping Speed) | Standard deviation of tapping speed in three repetitions. |
| {TT} | Mean (ICA) | Mean of Integrated cognitive assessment while tapping. |
| {TT} | SD (ICA) | Standard deviation of ICA. |
| {TT} | Mean (Error) | Mean of Error. |

5.6.2 Classification results

Table 5.3, Table 5.4, and Table 5.5 summarize the results of the RF and LR classifiers trained with the optimal features selected by MRMR and RFE-SVM, for the binary classification problem (young vs old adults) and the multi-class problem (5 age groups), respectively. Results in the following sections highlight the best classifier for each motor test and are grouped by hand.

Table 5.3: Classification accuracy (%) and AUCs of the RF and LR classifiers with the MRMR and RFE-SVM feature selection methods for motor performance classification (young vs. old adults)

| Task Name | Classifier | Feature Selection | Dominant Hand | | Non-Dominant Hand | |
|-----------|---------------------|-------------------|---------------|------|-------------------|------|
| | | | Accuracy | AUC | Accuracy | AUC |
| Circle A | RandomForest | RFE SVM | 86.0 | 0.70 | 86.4 | 0.67 |
| | | MRMR | 85.4 | 0.65 | 85.7 | 0.63 |
| | Logistic Regression | RFE SVM | 70.8 | 0.71 | 72.6 | 0.73 |
| | | MRMR | 62.9 | 0.66 | 66.2 | 0.67 |
| Square | RandomForest | RFE SVM | 85.5 | 0.65 | 85.9 | 0.66 |
| | | MRMR | 85.6 | 0.69 | 85.9 | 0.67 |
| | Logistic Regression | RFE SVM | 68.4 | 0.69 | 69.5 | 0.71 |
| | | MRMR | 64.2 | 0.67 | 63.7 | 0.69 |
| Circle- S | RandomForest | RFE SVM | 76.2 | 0.68 | 77.2 | 0.72 |
| | | MRMR | 77.2 | 0.69 | 77.0 | 0.72 |
| | Logistic Regression | RFE SVM | 68.6 | 0.71 | 70.4 | 0.74 |
| | | MRMR | 69.0 | 0.73 | 71.5 | 0.73 |
| Path | RandomForest | RFE SVM | 84.6 | 0.68 | 84.9 | 0.69 |
| | | MRMR | 84.2 | 0.68 | 85.6 | 0.69 |
| | Logistic Regression | RFE SVM | 66.4 | 0.72 | 67.9 | 0.73 |
| | | MRMR | 66.8 | 0.70 | 68.8 | 0.72 |
| Tapping-2 | RandomForest | RFE SVM | 80.2 | 0.55 | 81.0 | 0.58 |
| | | MRMR | 80.9 | 0.57 | 81.6 | 0.56 |
| | Logistic Regression | RFE SVM | 60.4 | 0.56 | 69.9 | 0.60 |
| | | MRMR | 68.2 | 0.56 | 69.8 | 0.63 |
| Tapping-4 | RandomForest | RFE SVM | 80.2 | 0.56 | 80.7 | 0.55 |
| | | MRMR | 79.3 | 0.57 | 79.6 | 0.55 |
| | Logistic Regression | RFE SVM | 61.9 | 0.58 | 65.9 | 0.56 |
| | | MRMR | 64.0 | 0.56 | 68.1 | 0.59 |

Young vs. old adults classification (dominant hand)

For the Circle-A motor test, the best-performing model was the RF classifier trained with 12 features selected from RFE-SVM, which achieved an accuracy of 86% and an AUC of 0.7.

For Square, the best-performing model was the RF classifier trained with 12 features selected from MRMR, which achieved an accuracy of 86% and an AUC of 0.69.

For Circle-S, the best model was the RF classifier trained with 12 features selected from MRMR, which achieved 77.2% accuracy, while the best AUC (0.73) was found for the LR classifier.

For Path, the best model was the RF classifier trained with 12 features from RFE-SVM, which achieved 84.6% accuracy, while the best AUC (0.72) was found for LR.

For Tapping-2, the best model was the RF classifier trained with 6 features from MRMR, which achieved an accuracy of 80.9% with an AUC of 0.57.

For Tapping-4, the best model was the RF classifier trained with 6 features from RFE-SVM, which achieved an accuracy of 80.2%, while the best AUC was obtained for

LR with 0.58.

Young vs. old adults classification (non-dominant hand)

For Circle-A, the best-performing model was the RF classifier trained with 12 features selected from RFE-SVM, which achieved an accuracy of 86.4% while the best AUC was achieved by LR (0.73).

For Square, the best-performing model was the RF classifier trained with 12 features selected from RFE-SVM, which achieved an accuracy of 85.9% while the best AUC was achieved by LR (0.71).

For Circle-S, the best-performing model was the RF classifier trained with 12 features selected from RFE-SVM, which achieved an accuracy of 77.2% while the best AUC was achieved by LR (0.74).

For Path, the best model was the RF classifier trained with 12 features selected from MRMR, which achieved 85.6% accuracy, while the best AUC of LR was 0.73.

For Tapping-2, the best model was the RF classifier trained with 6 features from MRMR, which achieved an accuracy of 81.6%, while the best AUC was achieved by LR (0.63).

For Tapping-4, the best model was the RF classifier trained with 6 features from RFE-SVM, which achieved an accuracy of 80.7%, while the best AUC was obtained by LR (0.59).

5 age groups classification

For Circle-A, the mean AUC of the two feature selection methods are 0.78 and 0.775 with RF and 0.64 and 0.645 with LR, for the dominant and non-dominant hands, respectively.

For Square, the mean AUC of the two feature selection methods are 0.78 and 0.79 with RF and 0.64 and 0.645 with LR, for the dominant and non-dominant hands, respectively.

For Circle-S, the mean AUC of the two feature selection methods are 0.7 and 0.715 with RF and 0.635 and 0.64 with LR, for the dominant and non-dominant hands, respectively.

For Path, the mean AUC of the two feature selection methods are 0.78 and 0.78 with RF and 0.645 and 0.655 with LR, for the dominant and non-dominant hands, respectively.

For Tapping-2, the mean AUC of the two feature selection methods are 0.72 and 0.72 with RF and 0.555 and 0.58 with LR, for the dominant and non-dominant hands, respectively.

For Tapping-4, the mean AUC of the two feature selection methods are: 0.72 and 0.71 with RF and 0.55 and 0.51 with LR, for the dominant and non-dominant hands, respectively.

Table 5.4: AUCs of the RF and LR classifiers with the MRMR and RFE-SVM feature selection methods for Circle-A, Square, and Circle-S motor performance classification (5 Age groups): a and b refer to dominant and non-dominant hand results.

| (a) | Feature Selection | Circle-A | Square | Circle-S |
|---|-------------------|--|--|--|
| Random Forest | MRMR | 0.78 (0.60, 0.52, 0.53, 0.60, 0.62) | 0.78 (0.62, 0.51, 0.54, 0.60, 0.65) | 0.70 (0.63, 0.52, 0.51, 0.55, 0.64) |
| | RFE_SVM | 0.78 (0.59, 0.51, 0.51, 0.59, 0.68) | 0.78 (0.61, 0.53, 0.52, 0.59, 0.64) | 0.70 (0.61, 0.51, 0.49, 0.56, 0.63) |
| Logistic Regression | MRMR | 0.61 (0.59, 0.53, 0.53, 0.59, 0.63) | 0.62 (0.62, 0.52, 0.55, 0.58, 0.63) | 0.64 (0.66, 0.50, 0.50, 0.63, 0.69) |
| | RFE_SVM | 0.66 (0.62, 0.50, 0.51, 0.63, 0.69) | 0.66 (0.64, 0.51, 0.56, 0.63, 0.67) | 0.63 (0.65, 0.50, 0.52, 0.63, 0.70) |
| (b) | | | | |
| Random Forest | MRMR | 0.77 (0.58, 0.51, 0.53, 0.57, 0.60) | 0.79 (0.63, 0.53, 0.57, 0.60, 0.63) | 0.72 (0.66, 0.51, 0.55, 0.62, 0.65) |
| | RFE_SVM | 0.78 (0.60, 0.52, 0.54, 0.58, 0.60) | 0.79 (0.64, 0.50, 0.58, 0.61, 0.63) | 0.71 (0.65, 0.51, 0.56, 0.61, 0.64) |
| Logistic Regression | MRMR | 0.63 (0.61, 0.52, 0.51, 0.58, 0.66) | 0.63 (0.64, 0.53, 0.59, 0.59, 0.62) | 0.65 (0.68, 0.48, 0.54, 0.61, 0.70) |
| | RFE_SVM | 0.66 (0.62, 0.51, 0.52, 0.62, 0.67) | 0.66 (0.65, 0.52, 0.59, 0.64, 0.69) | 0.63 (0.65, 0.47, 0.54, 0.62, 0.68) |
| Note: microAUCs of five age groups is displayed in an order (Group-A, B, C, D, and E) below the weighted average AUC. | | | | |

Table 5.5: AUCs of the RF and LR classifiers with the MRMR and RFE-SVM feature selection methods for Path, Tapping-2, and Tapping-4 motor performance classification (5 Age groups): a and b refer to dominant and non-dominant hand results.

| (a) | Feature Selection | Path | Tapping-2 | Tapping-4 |
|---|-------------------|--|--|--|
| Random Forest | MRMR | 0.78 (0.62, 0.53, 0.53, 0.58, 0.72) | 0.72 (0.53, 0.50, 0.52, 0.54, 0.53) | 0.73 (0.54, 0.50, 0.51, 0.53, 0.53) |
| | RFE_SVM | 0.78 (0.61, 0.52, 0.51, 0.57, 0.70) | 0.72 (0.51, 0.51, 0.53, 0.54, 0.56) | 0.71 (0.53, 0.52, 0.51, 0.52, 0.57) |
| Logistic Regression | MRMR | 0.64 (0.61, 0.52, 0.51, 0.61, 0.73) | 0.56 (0.53, 0.50, 0.49, 0.53, 0.56) | 0.55 (0.53, 0.49, 0.51, 0.51, 0.53) |
| | RFE_SVM | 0.65 (0.63, 0.52, 0.52, 0.61, 0.73) | 0.55 (0.53, 0.50, 0.48, 0.52, 0.52) | 0.55 (0.53, 0.49, 0.50, 0.51, 0.55) |
| (b) | | | | |
| Random Forest | MRMR | 0.78 (0.64, 0.53, 0.55, 0.59, 0.67) | 0.72 (0.52, 0.48, 0.52, 0.53, 0.50) | 0.72 (0.51, 0.50, 0.49, 0.50, 0.47) |
| | RFE_SVM | 0.78 (0.63, 0.53, 0.54, 0.59, 0.68) | 0.72 (0.52, 0.51, 0.51, 0.52, 0.54) | 0.7 (0.52, 0.50, 0.49, 0.49, 0.48) |
| Logistic Regression | MRMR | 0.66 (0.65, 0.50, 0.51, 0.61, 0.73) | 0.58 (0.55, 0.48, 0.51, 0.57, 0.55) | 0.5 (0.52, 0.49, 0.48, 0.50, 0.48) |
| | RFE_SVM | 0.65 (0.64, 0.50, 0.49, 0.63, 0.74) | 0.58 (0.54, 0.49, 0.50, 0.57, 0.53) | 0.52 (0.52, 0.52, 0.48, 0.47, 0.55) |
| Note: microAUCs of five age groups is displayed in an order (Group-A, B, C, D, and E) below the weighted average AUC. | | | | |

5.7 Discussion

The results of feature selection show that all the additional features we extracted from the data besides the performance measures we used in the statistical analysis were always included in the selected set, highlighting their usefulness in characterizing motor patterns [90, 51, 52, 85]. The overall number of features we considered was limited with respect to other studies in the literature and this could have affected the performance of our models. In the future, it would be interesting to identify more features, either manually or through automatic processes, that could add more discriminatory information to the process, such as features related to the shape of trajectories.

The results of the binary (young vs old) classification task, summarized in Table 5.3, show that the RF and LR classifiers achieved a good degree of predictability with both MRMR and RFE-SVM selected features, as the AUC values are always above 0.5 for all combinations, ranging from 0.55 to 0.74. The accuracy and AUCs values for the non-dominant hand data were often found to be higher than those for the dominant hand, especially with the LR classifier. In particular, the non-dominant hand models showed improvements in AUC values from 0.1 to 0.2 for the trial making tests and from

0.1 to 0.6 for the tapping tests, while accuracy values showed significant improvement for tapping tests (almost 10% for Tapping-2 with RFE-SVM selected features and the LR classifier). This shows that motor tests that are carried out with the non-dominant hand can provide more discriminative data compared to tests with the dominant hand. In terms of accuracy, the Random Forest classifier almost always showed significantly better results compared to the Logistic Regression classifier, for both the dominant and non-dominant hand. The difference between the two classifiers was on average around 20%, regardless of the feature selection method employed. The accuracy of the RF classifier was typically over 80%, except for the Circle-S test, where it reached 76%. Circle-A, Square, and Path were the motor test that achieved better results (around 85% accuracy and 0.7 AUC). Overall, performance of the RF classifier was good, both in terms of accuracy and AUC, even if it did not reach the performance of models available in the literature for the discrimination between healthy subjects and subjects with neurological disorders ([93, 183, 5, 116]). Further optimizations of the process and the inclusion of a larger set of initial features would probably lead to even better performance.

The results of the multiclass (5 age groups) classification task, summarized in Table 5.4 and Table 5.5, show mixed results, as was expected from the statistical analysis. AUC values for the different age groups were typically lower than what we found in the binary classification task, with values for some of the motor tasks being lower than 0.5. AUC values of Group E (the oldest group) are always the highest, followed by Group A (the youngest group) and Group D (the second oldest group) with similar results. Both Group-B and Group-C have the lowest AUCs. There are also significant differences among motor tests, with tapping tests showing significantly lower AUC values compared to the other tests, especially for Group A and Group E. The impact of the two feature selection methods seem to be minimal. Performance of the two classifiers does not seem to differ much in terms of group-based AUCs while the weighted average

AUC values are better for the RF classifier than the LR classifier. Overall, these results show that classification of the extreme age groups is much easier than classification of the middle age groups.

The inaccuracies in age group recognition may have different reasons. For example, results may be distorted by the fact that we collected the data remotely and users could have consciously or unconsciously provided an incorrect age. In the future, we will explore different features, classification algorithms (e.g., ensemble methods), and classification parameters to reach improved classification results.

PART II
**Computer-aided analysis of MRI datasets for
brain tumor survival prediction**

6

Background: brain tumor survival prediction with radiomics analysis

Predicting the survival of brain tumor patients is an important task for their treatment and surgical planning. One of the approaches that can be used to obtain such prediction is radiomics analysis of medical images of the tumors. The radiomics process consists of different stages that are applied in sequence to the input images and their derived data. In this part of the thesis, we specifically focus on two of the radiomics process stages: segmentation and feature selection. More specifically, we focus on evaluating the impact of different segmentation algorithms and of feature robustness on the prediction of the survival of patients affected by one specific type of brain tumors, High Grade Gliomas (HGGs).

In this chapter, we introduce gliomas, one of the most common types of brain tu-

mors, and describe the stages of the radiomics process that can be used to automatically analyze medical images of brain tumors. We then survey related work on segmentation algorithms and feature robustness, before presenting the state-of-the-art on the application of the full radiomics process to brain tumor survival prediction. We conclude the chapter by summarizing our research goals.

6.1 Gliomas

A brain tumor is a benign or malignant mass of abnormal cells that grows in the brain. Among the many types of brain tumors, gliomas are the most common. They arise from the glial cells that surround neurons [29]. Based on the classification scheme introduced by the World Health Organization (WHO) in 2016, gliomas can be divided into four classes [105]: grade I and grade II gliomas are called Low-Grade Gliomas (LGGs), are less aggressive, have a better prognosis, and are more common in children and young adults; grade III and grade IV gliomas are considered High-Grade Gliomas (HGGs), are malignant, have a worse prognosis, and are more common in older adults.

To give an idea of the impact of brain tumors in a western country, 23 out of 100,000 people were diagnosed with a brain tumor each year between 2011 and 2015 in the United States (See Figure 6.1) [122]. HGGs account for about 80% of malignant brain tumors. The prevalence of HGGs is 0.59 to 5 out of 100,000 people and the number of people diagnosed with this type of glioma is rising continually [64]. Different motivations have been identified for the increase in glioma diagnoses including the aging of the population, exposure to ionizing radiation, air pollution, but also the fact that there is now a wider access to medical imaging that makes discovery more frequent [64].

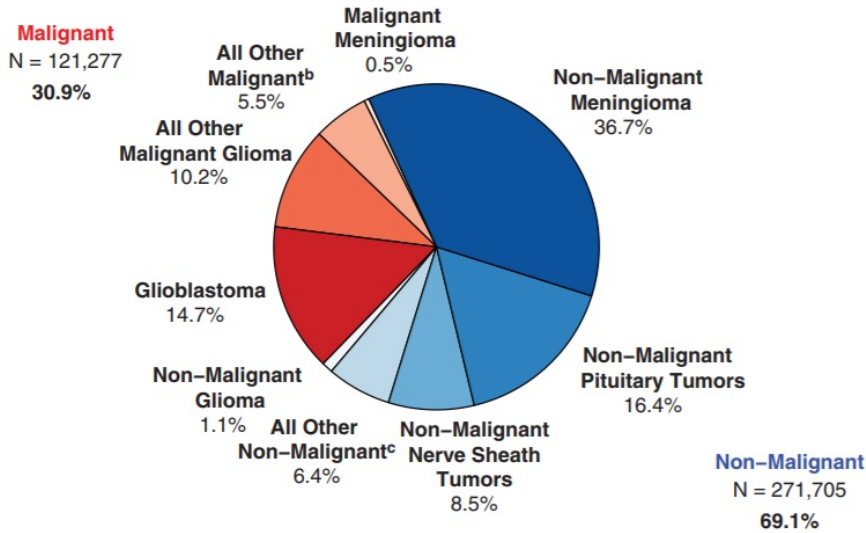


Figure 6.1: Distribution of brain tumors in 2011-2015 in the United States [122].

6.2 Medical imaging for brain tumor diagnosis and evaluation

Different types of medical imaging, such as Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Digital Pathology Images (DPI), and X-rays techniques, can be employed to perform diagnostic and prognostic evaluations of brain tumors. In the early stages of brain tumors, CT images can provide good diagnosis [173]. In later stages, 3D Multi-parametric MRI is a better solution for brain tumor detection and segmentation [115, 12, 129]. MRI is a non-invasive imaging modality that is routinely used for three-dimensional spatial localization of brain tumors. Unlike X-ray and CT imaging, MRI provides high resolution images, with superior soft tissue contrast without employing ionizing radiation [14]. For the diagnosis of brain gliomas, four MRI sequences are routinely acquired, which are called T1-weighted (T1), T1-weighted contrast enhanced (T1ce), T2 weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR).

These multi-parametric scans provide useful additional information for proper tumour sub-region delineation. Whole tumor and tumor core regions are highlighted by T2 and FLAIR MRI scans while enhancing tumor region and necrotic component of tumor core are highlighted by T1 and T1ce MRI scans.

6.3 Overall survival prediction

Overall survival (OS) is defined as the number of days a patient survives post-surgery [109]. Patient survival is strongly related to tumor grade. LGGs are less aggressive and have a survival time of several years, while HGGs are more aggressive and malignant with a median survival time of 12 – 16 months, even after good treatment [20]. With the passage of time, LGG gliomas can grow and change into HGG. HGGs are heterogeneous in nature and lead to shorter survival time due to rapid tumor growth and tumor invasion into surrounding brain tissues [83]. In the management of HGGs, OS plays an important role for treatment and surgical planning [147, 55, 136].

Different approaches can be used to estimate OS. In the clinical approach, survival of HGGs patients is predicted with clinical characteristics such as patient’s age, gender, performance, and resection status and with pathological characteristics of the tumor such as WHO grade and morphology [153]. However, this approach has limitations because gliomas are heterogeneous in nature and may have different locations, sizes, and tumor-affected regions that make prediction more difficult.

Another approach that has become increasingly common in recent years is to predict OS of brain gliomas by directly using information extracted from the medical images of the tumors through *radiomics analysis*. Radiomics is an emerging method that extracts meaningful high-dimensional data from medical images into meaningful and quantitative predictive data with high precision and throughput. Extracted features have the ability to exhibit the distinct phenotypic characteristics of tumor patterns that would not be

noticed by the human eye, thus enhancing the clinical significance and the prognostic and predictive power for HGG patients. Various imaging modalities such as MRI, PET and CT can be used for radiomics analysis. At the beginning of the analysis, the most appropriate imaging modalities for the specific clinical questions to answer must be identified [123]. Because of its advantages over other modalities in the context of brain tumor diagnosis and evaluation, MRI is often the source of choice for radiomics analysis aimed at addressing different clinical questions, such as treatment response [38], tumor classification [171], disease progression [140], tumor grading [128], and survival prediction [178, 55].

In the radiomics-based approach, OS prediction for HGGs can be formulated as a classification task. The Brain Tumor Segmentation (BraTS) challenge, which focuses on the evaluation of state-of-the-art methods for the segmentation of brain tumors in multimodal MRI scans and on the prediction of patient overall survival, defines three survival classes: short-term survival (< 10 months), medium-term survival ($10 - 15$ months), and long-term (> 15 months) survival. Because of its impact on the field, the publicly available BraTS dataset [115, 11, 10, 12] has become a de-facto standard for brain tumor segmentation and OS classification, facilitating rigorous comparison of available methods and leading to substantial advances in the field.

6.4 The radiomics process

The radiomics process consists of several sub-processes with defined inputs and outputs:

1. Preprocessing of the medical images prior to analysis
2. Segmentation of the Region Of Interest (ROI) in the images into different sub-regions
3. Extraction of radiomic features from the segmented regions

4. Selection of the most predictive set of features
5. Development of a machine learning model based on the selected radiomic features and use of the model for the target task
6. Evaluation of the results based on performance metrics

6.4.1 Preprocessing

The medical images to analyze are often acquired by different institutions using different imaging protocols. To standardize the images, some preprocessing steps must be performed before any analysis can be done. Bias field correction is a recommended preprocessing step to compensate for low-frequency changes in the images caused by field inhomogeneities [164]. In addition, intensity normalization is used to convert the image signal intensity to a standard intensity range. The intensity distribution is adjusted by z-score normalization (zero mean and unit variance). To benefit from multi-parametric MRI sequences, all sequences must be registered (aligned). The registration step involves co-registration of all modalities or registration to an atlas space. MRI registration to an atlas with linear registration creates a common reference frame. The MRI scans are also skull-stripped and resampled to an isotropic resolution of $1 \times 1 \times 1 \text{ mm}^3$ voxels. The preprocessing of multi-parametric MRI scans before using them for automatic segmentation can reduce the machine learning training time and increase model performance.

6.4.2 Segmentation

Segmentation of the region of interest (ROI) in the medical images is a necessary step for feature extraction. Precise and accurate tumor segmentation is critical for treatment and surgical planning, tumor characterization and patient survival prediction [59]. Manual segmentation of the ROI is considered the gold standard, but it is not a viable solution for large datasets because it is subject to some degree of bias. Radiomics-based

approaches thus often use automatic segmentation algorithms. Many automatic segmentation algorithms have been developed for segmenting the ROI. Among these, Deep Learning (DL) algorithms based on Convolutional Neural Networks (CNNs) have recently emerged as a promising solution for automatic segmentation of tumor subregions based on multi-parametric MRI scans. A recent review by Liu et.al. [104] provides a comprehensive review of DL-based segmentation algorithms for brain tumors and reports that accurate segmentation of tumor regions from MRI is still a challenging task. A variety of tools are also available to support the segmentation task, such as GLISTR [63], BraTuMIA [114, 134], 3D Slicer [54] and ITK-SNAP¹.

6.4.3 Feature extraction

Based on the segmentation results, features are extracted from the segmented regions. Radiomic features can be divided into agnostic features and semantic features [62]. Semantic features used by radiologists include lesion volume, diameter, and morphology. Agnostic features are mathematically-derived quantitative features.

These features capture imaging patterns such as:

- First-order statistical features calculated from the voxels within the ROI, including mean, median, standard deviation, range, and entropy.
- Shape-based features that describe the geometric shape of the tumor in the 3D surface, including volume, diameter, and surface area of the tumor, as well as some derived values such as sphericity, compactness, spherical disproportion, and flatness.
- Texture features that account for image contrast between voxels in spatial relationships, including the grey-level run-length matrix, the grey-level co-occurrence matrix, and the grey-level size zone matrix.

¹<http://www.itksnap.org>.

Many different tools and libraries exist to support feature extraction. Examples include CaPTk [41], PyRadiomics [165], simple ITK filters [107] and some MATLAB based tools (The Mathworks, Natick, MA). The implementation of these features may vary, but the standard definitions for all of them are provided by the Image biomarker standardisation initiative (IBSI) [187].

6.4.4 Feature selection

Radiomics projects usually extract a larger number of features than the number of samples. A machine learning model built with such a number of features will result in overfitting. The extracted set of radiomics features may also contain some redundant and non-informative features. Therefore, dimension reduction and selection of the k most robust, task-specific, and discriminatory features are important steps before training a machine learning model. The possible stages for feature selection are outlined in Figure 6.2.

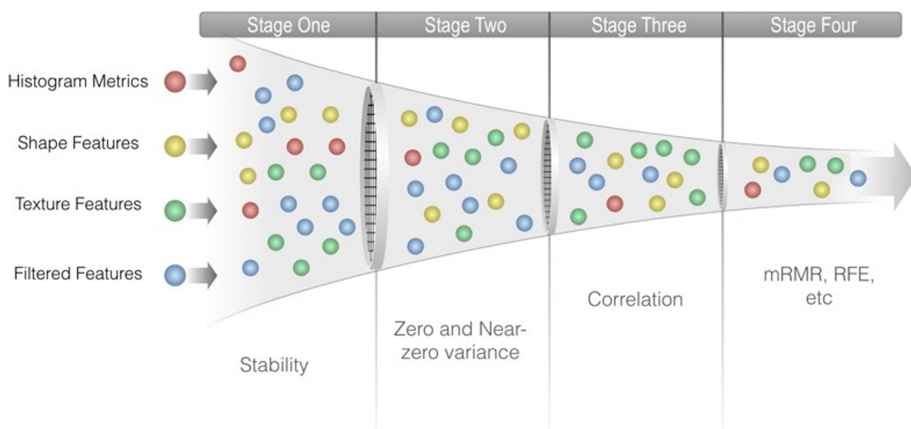


Figure 6.2: Feature selection is a multi-stage process consisting of removing unstable features, removing features with zero or near zero variance (non-informative features), removing highly correlated features (redundant features), selecting the optimal set of feature by using wrapper methods like maximum relevance minimum redundancy (mRMR) or recursive feature elimination (RFE) [123].

Feature selection involves two main steps: *feature set reduction* and *feature subset selection*. In *feature set reduction*, a reasonable set of features is selected from a large pool of features using statistical approaches. Feature stability can be achieved by avoiding features that are sensitive to scanner, acquisition parameters, reconstruction algorithm, and segmentation variability [91]. Features with low dynamic range are not informative and are eliminated with a mean absolute deviation (MAD) of zero. Highly correlated features increase the likelihood of overfitting and are therefore removed using task-specific reduction methods such as the Spearman correlation coefficient and concordance index. At the end of the feature reduction step, we obtain a set of informative, stable, and non-redundant features. However, this set may still contain too many features for training the machine learning model.

The optimal and most predictive set of features is then identified by the *feature subset selection* step. Feature subset selection methods can be grouped into three categories [158]:

1. Filtering methods select the most informative set of features using statistical properties of the data such as feature correlation, feature consistency, and information theoretic measures. These methods are computationally efficient because they do not use a learning algorithm.
2. Embedded methods use an optimization approach for feature selection that works in combination with machine learning models such as LASSO [58], Ridge Regression [179], and Elastic Net [159]. These methods are computationally intensive as they depend on the trained classifier.
3. Wrapper methods search for a subset of relevant and non-redundant features and then evaluate these subsets based on the performance of the already selected classifier until an optimal set of features is determined. Examples include brute force, where all possible feature combinations are tested using a machine learning model,

and Recursive feature elimination with cross-validation (RFE-CV).

6.4.5 Machine learning model development

A machine learning model is built with the most predictive features for a given clinical question. Depending on the nature of the clinical question, regression or classification methods are used and the result will be in discrete or continuous form. The available data will ideally be divided into a training, validation, and testing cohort. One large cohort will be used for training the model and fine-tuning the hyperparameters, and a smaller validation cohort will be used for validating the model performance. Finally, the trained model is applied to the test cohort.

In the case of a small data set, the K-fold cross-validation technique can be used. The original data set is divided into k subsets, one of which is used for validation while the other $k-1$ subsets are used for training the model. The recommended value of k , as found in various studies, is 5 or 10 [147, 88].

As is common in other contexts, machine-learning methods for radiomics analysis can be supervised, where the available data consists of labelled examples, or unsupervised, where patterns are learned automatically from untagged data.

6.4.6 Evaluation

The final step in the radiomics process is to evaluate the performance of the process. Different performance measures are used based on the target application. In most cases, regression and classification tasks are evaluated using an accuracy metric. Classification performance has been evaluated using various metrics such as accuracy, sensitivity, specificity, and precision. To determine the overall classification performance, the area under the receiver operating curve (ROC-AUC), balanced accuracy, F1 score, and Area Under the precision recall curves (AUPRC) can be used. For unbalanced data sets, the

AUC and AUPRC are more important than other measures. Confusion matrices help to understand the classifier’s per class performance.

6.5 State of the art on brain tumor segmentation

Brain tumor volume can be partitioned into three non-overlapping or three overlapping subregions. The three non-overlapping subregions of brain tumor volume are called peritumoral edema (PTE), non-enhancing core (NEC), and enhancing core (ENC) (see Figure 6.3). These non-overlapping subregions can be combined in various ways to generate three overlapping subregions of brain tumor volume called Whole Tumor (WT), Tumor Core (TC), and Enhancing Tumor (EC). WT is a combination of the PTE, NEC, and ENC subregions. TC is a combination of NEC and ENC. ET only contains the enhancing core (ENC).

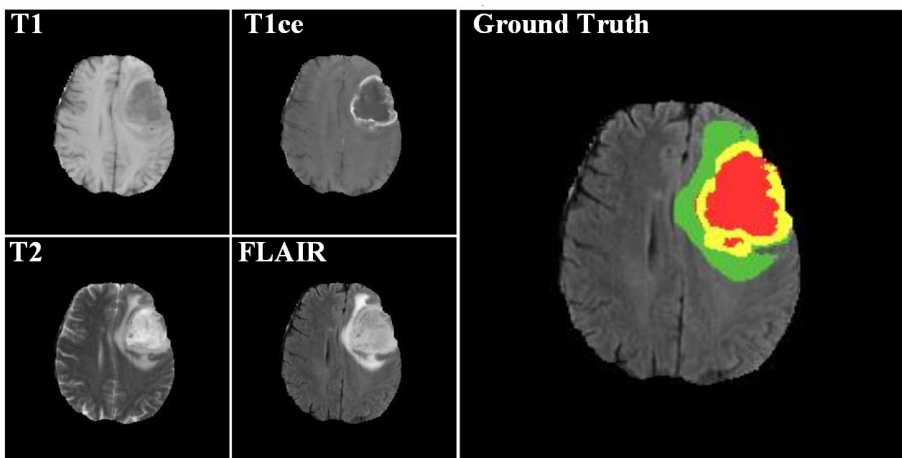


Figure 6.3: Example input dataset with four MRI modalities and corresponding ground truth segmentation map. The last frame on the right is the ground truth with corresponding manual segmentation annotation. Label legend: enhancing tumor (green), peritumoral edema (yellow) and necrotic and non-enhancing tumor (red) [104].

While manual segmentation by experienced radiologists is the gold standard for segmenting a volume of interest, such approach suffers from inter-reader variability of

74% – 85% when large datasets are taken into consideration [115]. Moreover, manual segmentation has very low reproducibility. The use of automatic segmentation algorithms does not suffer from these problems. However, automatic segmentation of brain tumors and their subregions is a challenging task for a number of reasons including heterogeneity of tumor shape and appearance, ambiguous tumour boundaries, lack of high-quality imaging data, unbalanced tumor tissue, presence of artifacts, and high computational and storage requirements due to the volumetric nature of the data and processing requirements.

Various automatic segmentation approaches have been proposed in the literature. These include the thresholding method [182], in which voxels above a threshold are classified as belonging to the tumor, the edge-based method [6], in which changes in the intensity between edges of voxels are used as boundaries of tumors, the region growing method [101], in which voxels that are similar to a seed voxel are classified as belonging to the tumor, Convolutional Neural Networks (CNNs) [32, 46, 157, 108, 75, 170], which are based on the structure of the human visual cortex, and the Atlas method [17], in which a tumor-free reference MRI is used to segment the MRI containing the tumor volume.

Many of these automatic segmentation methods have demonstrated their effectiveness in segmenting brain tumors and their sub-regions. However, the results of each method are different and there is no single method that provides precise segmentation. Each segmentation algorithm works differently, and even the same algorithm gives different results with different parameters.

In the following sections, we provide details on notable segmentation algorithms that have been proposed in the literature and that have been found to perform well in segmenting tumor subregions on the BraTS datasets.

6.5.1 CNN-based segmentation of brain tumor volume

In recent years, CNNs have shown promising performance in segmenting medical images. Among the various CNN architectures, U-Net [143] and its variants [32, 46, 157, 108, 75, 170] proved their efficiency for medical image segmentation. A typical U-Net is an encoder-decoder architecture. The encoder is used to map the input image into latent space, followed by a decoder path. Encoder-decoder branches at the same depth level are connected by skip connections to obtain the high level spatial information. Each layer consists of convolutions followed by rectified linear units (ReLU) and max-pooling operations only for the down-sampling path. This network does not have a fully connected layer since it is a fully convolutional network. At the end, a softmax activation layer is used to map the input to the target segmentation labels.

DeepMedicRes 3D CNN

Kamnitsas et al [81] have proposed a carefully designed 3D CNN called DeepMedic that not only uses small kernels, but also modifies the receptive field of the network to enable dense inference. The architecture includes two paths that process segments at two different scales, allowing the use of additional contextual information. They proposed to sample training segments centered on a foreground voxel (voxel representing the tumor region) and a background voxel (voxel representing the healthy brain region) with a probability of 50% to correct the class imbalance inherent in the dataset, and they showed that this corrects the true class distribution of the dataset. Finally, to refine the boundary between the whole tumor and the background and reduce isolated false positives, the probability maps obtained at the output were post-processed with a 3D Conditional Random Field. DeepMedicRes 3D CNN provides non-overlapping subregions (PTE, NEC, and ENC) of brain tumor volume.

Dong 2D U-Net

Dong et al. [46] proposed a slice-based 2D U-Net to segment tumor sub-regions on volumetric MRI scans. Instead of four available MRI sequences, it uses only two sequences highlighting different tumor subregions. The authors suggested using FLAIR images to delineate the whole tumor and tumor core regions and T1ce images to segment the enhanced core region. Using a minimal number of MRI scans per patient also reduces computational costs. To account for variations in tumor size and appearance, they used an extensive data augmentation pipeline. Dong 2D U-Net provides non-overlapping subregions (PTE, NEC, and ENC) of brain tumor volume.

Wang 2.5D CNN

Wang et al. [170] developed a hierarchical region-based approach to tumor segmentation, in which the multiclass segmentation problem is split into a hierarchical three-stage binary segmentation task. First, the whole tumor is segmented by a 2.5D CNN, the result of which is forwarded to another 2.5D CNN network that segments the tumor core within the whole tumor. The third 2.5D CNN receives the segmentation of the tumor core and segments enhancing tumor core within the tumor core. The outputs of the three 2.5D CNNs are combined to produce the final segmentation map. The three networks are trained independently and form a cascade during segmentation, a technique called region-based prediction. Each CNN uses anisotropic pseudo-3D convolutional kernels with multiscale prediction and uses fusion of multiple views (in axial, coronal, and sagittal directions) to create the segmentation of a tumor subregion. Three such segmentation frameworks, namely W-Net, T-Net, and E-Net, were combined in a cascade where one framework operates on the output of the previous one to create a multiclass segmentation map. Wang 2.5D U-Net yields overlapping subregions (WT, TC, and EC) of a brain tumor.

Isensee 3D U-Net

Isensee et al. [75] made the argument that instead of proposing different architectures, one should focus on the training procedure, hence the title of their paper ‘No New-Net’. Their premise is that a carefully tuned and well-trained U-Net is hard to beat in segmentation performance and, hence, instead of focusing on architectural modifications, one should emphasize the training and testing procedures. Their approach differed from other previously proposed methods in several ways. Instead of the more prevalent Batch Normalization, they employed Instance Normalization which was more consistent with their small batch size. An un-weighted combination (sum) of Cross Entropy and Dice loss was used to address poorly calibrated softmax probabilities, and occasional convergence issues due to high variance in using Dice loss alone. They introduced co-training with two combinations of three datasets which is equivalent to training on an augmented dataset. Lastly, they make use of region-based prediction inspired by Wang et al. [170]. The networks are trained with 5-fold cross-validation and each network in the hierarchical cascade is an ensemble of five network models, one from each fold. This amounts to employing thirty different networks to segment a patient scan. Isensee 3D U-Net yields overlapping subregions (WT, TC, and EC) of a brain tumor.

Pereira 2D U-Net

In their patch wise segmentation, Pereira et al. [131] propose a 2D network that predicts the class for the central voxel in a 33x33 patch. They investigated small 3x3 kernels, which allow to design a deeper architecture with few weights and act as an implicit regularizer against overfitting. They also proposed two different networks: a shallow design for LGG segmentation and a deeper architecture for HGG segmentation. The entire segmentation task is divided into three steps: preprocessing multiparametric MRI scans, training a CNN, and postprocessing the predicted segmentation maps.

Preprocessing includes N4 bias correction, Nyul intensity normalization, and z-score normalization. Using a data augmentation technique that accounts for the anatomical heterogeneity of brain tumors and provides better generalization performance, network training is performed for the extracted patches. A morphological closure and cluster thresholding were used as post-processing steps for the segmentations generated by the trained network at the test time. By removing erroneous segmented clusters, this step improves the segmentation performance. Pereira 2D U-Net provides non-overlapping subregions (PTE, NEC, and ENC) of brain tumor volume.

HDC-Net

Luo et al. [108] proposed a hierarchically decoupled CNN (HDC-Net) by replacing standard convolution blocks in a 3D U-Net with new lightweight HDC blocks composed of carefully arranged 2D convolutions. HDC blocks have low computational complexity and work simultaneously for the channel and spatial dimensions. For the spatial domain, view decoupled works in three views (i.e. axial, coronal, sagittal). For the channel domain, 2D convolution is only applied on the axial view, its view is applied on other feature channels. HDC-Net provides non-overlapping subregions (PTE, NEC, and ENC) of brain tumor volume.

E_1D_3 3D U-Net

Talha Bukhari and Mohy-ud-Din [157] proposed a modification of the 3D U-Net inspired by the concept of TreeNets and region-based prediction. The proposal, named E_1D_3 , is a single-encoder-multi-decoder architecture in which each decoder segments one of the three hierarchical tumor subregions: WT, TC, and EC. The three binary segmentation maps are then merged through a combination of morphological processing, cluster thresholding, and hierarchical operations to create a multi-class segmentation map. E_1D_3 3D U-Net yields overlapping subregions (WT, TC, and EC) of the brain

tumor volume.

Comparison of segmentation algorithms

When comparing the effectiveness of the above-mentioned algorithms in segmenting brain tumors and their sub-regions, it is clear that no single method provides precise segmentation in all contexts. For example, Pereira 2D U-Net has shown superior performance on the BraTS 2013 dataset, Dong 2D U-Net has been found to have superior performance on the BraTS 2015 dataset, DeepMedicRes on the BraTS 2016 dataset, Wang 2.5D CNN on the BraTS 2017 and 2018 datasets, Isensee 3D U-Net on the BraTS 2018 dataset, HDC-Net on the BraTS 2017 dataset, and E_1D_3 3D U-Net on the BraTS 2018 and 2021 datasets. For comparison purposes, Table 6.1 shows performance results of several state-of-the-art algorithms in segmenting tumor subregions on the BraTS validation dataset from 2018 to 2021.

6.5.2 Atlas-based segmentation

Image registration is a process used to align multimodal MR images to a common frame of reference. Each tumor is located at a different site in the brain, which can be determined by registering the patient’s MR image with a predefined anatomical atlas. In medical image analysis, registration is an essential step for image fusion, voxel-based analysis and image segmentation. In atlas based segmentation, an atlas with predefined labels is registered into a target image, and segmentation is obtained by overlaying the expert annotated region of interest with the registered atlas. Registration of glioma MR images suffers from two main problems. First, tumor components produce an alteration of the contrast of normal tissue. Second, the shape and volume of normal tissue change due to surgery and tumor volume [168].

There are several approaches to image registration that differ in terms of degrees of freedom, optimization choices, and similarity metrics. Two types of registration schemes

Table 6.1: Comparison of the state-of-the-art methods on the BraTS validation dataset (years 2018-2021).

| Segmentation Network | Dice Similarity Coefficient (%) | | | Hausdorff Distance (mm) | | |
|-------------------------------|---------------------------------|------|------|-------------------------|------|-------|
| | WT | TC | EC | WT | TC | EC |
| Self Ensembled U-Net [186] | 91.0 | 84.0 | 75.0 | 4.57 | 5.58 | 3.84 |
| Scale Attention Network [181] | 91.0 | 85.0 | 79.0 | 4.09 | 5.88 | 18.19 |
| Self-ensembled 3D U-Net [68] | 91.0 | 85.0 | 80.0 | 4.30 | 5.69 | 20.56 |
| Patch based N-Net [84] | 90.5 | 81.3 | 78.8 | 4.32 | 7.56 | 3.81 |
| Cascaded V-Net [70] | 90.5 | 83.6 | 77.7 | 5.18 | 6.28 | 3.51 |
| 3D-SE Inception [180] | 90.1 | 81.3 | 79.8 | 6.37 | 8.84 | 4.16 |
| Cross-Modality GAN [184] | 90.3 | 83.6 | 79.1 | 5.00 | 6.37 | 3.99 |
| Cascaded-Attention-Net [175] | 90.1 | 82.6 | 78.5 | 6.39 | 6.28 | 3.81 |
| E1D3 3D U-Net [157] | 91.2 | 85.7 | 80.7 | 6.11 | 5.54 | 3.12 |
| Isensee 3D U-Net [75] | 90.9 | 85.2 | 80.7 | 5.83 | 7.20 | 2.74 |
| HDC-Net [108] | 89.7 | 84.7 | 80.9 | 4.12 | 6.12 | 2.43 |

are possible, linear and non-linear. Linear registration is a simple and most commonly used type of registration. It is based on a rigid transformation with 6 parameters (rotation and translation on the x, y and z axis) and an affine transformation with 12 parameters (rotation, translation, scaling and shear on the x, y and z axis). This type of registration works globally. Non-linear registration provides local matching of tissues in the brain to the template. Generally, non-linear registration is initialized with the result of linear registration.

Registration of gliomas is challenging, as it requires to register the brain image with gliomas into a healthy template image. Therefore, it is important to ensure that intensity

differences within tumor regions are excluded when calculating the cost function for good quality registration.

6.6 State-of-the-art on robustness of radiomic features

Radiomic approaches extract quantitative features from radiological images and enable diagnostic and prognostic evaluation of various diseases. However, these radiological images are acquired under different acquisition protocols and preprocessed using different algorithms. The features extracted from these images are susceptible to these variations. The proper translation of radiomic models into clinical setting requires fully automated and generalized models that must include features that are robust, i.e. stable, with respect to these variations.

The reliability of radiomic features can vary depending on the uncertainty of tumor segmentation, as features are extracted from the obtained tumor segmentation maps. Most studies have focused on semi-automatic and inter-rater segmentation variability, but fully automatic segmentation variability has not yet been demonstrated in radiomics.

Tixier et al [161] investigated brain tumor segmentation variability between semi-automatic and inter-rater manual segmentations by two raters for MR images. They reported that GLCM texture features were robust and the subset of GLSZM features was robust for interactive manual segmentation. They found that variations in most radiomic features were greater between two consecutive scans than between segmentations.

Haarburger et al [67] analyzed the reproducibility of radiomic features extracted from manual segmentations confirmed by four experienced raters and from probabilistic segmentations obtained by probabilistic U-Net. They found that first-order and shape features are robust to segmentation variability.

Kalpathy-cramer et al [80] evaluated variations in radiomic features in automatic

segmentation of lung nodules and different feature implementations. They reported that 68% of 830 features had an overall correlation coefficient (OCCC) ≥ 0.75 .

Suter et al. [156] studied the robustness of various feature categories using 125 perturbations including varying image resolution, k-space subsampling, additive noise, and bin width for gray values. The study showed that, for the OS classification task, shape features are most robust, with intra-class correlation coefficient (ICC) $\in [0.97, 0.99]$, followed by first order features, with $ICC \in [0.48, 0.92]$, texture features such as GLSZM with $ICC \in [0.28, 0.83]$, GLCM with $ICC \in [0.32, 0.82]$, GLRLM with $ICC \in [0.30, 0.80]$, GLDM with $ICC \in [0.31, 0.78]$, and deep features with $ICC \in [0.48, 0.86]$.

6.7 State of the art on radiomics-based OS prediction

Many different radiomics-based solutions have been proposed in the literature for OS prediction, mostly in connection with the BraTS challenge. In the following, we will present the most significant works that have been published in recent years.

Puybureau et al. [136] used a 2D fully convolutional neural network (FCNN), based on the VGG-16 architecture, for segmentation of brain tumors into three non-overlapping subregions: peritumoral edema (PTE), non-enhancing core (NEC), and enhancing core (ENC). Ten volumetric features were extracted from scans of subjects with Gross Tumor Resection (GTR) status, using ground truth segmentation maps on the training cohort and the obtained multi-regional segmentation maps on the validation and challenge cohorts. The extracted features were normalized with principal component analysis and used to train 50 Random Forest classifiers. The final prediction (of survival class) was obtained by a majority voting on the 50 predictions from trained classifiers. The authors reported an accuracy of 37.9% on the validation cohort and 61% on the challenge cohort.

Kao et al. [84] utilized an ensemble of twenty-six neural network architectures (nineteen variants of Deep-Medic [82] and seven variants of 3D U-Net [32] with random

initialization, data augmentation, normalization, and loss function) for segmentation of brain tumor into three non-overlapping subregions (PTE, NEC, and ENC). From the obtained multi-regional segmentation maps they extracted 19 morphological, 19 volumetric, 78 volumetric spatial and 116 tractography features from 59 subjects with Gross Tumor Resection (GTR) status. Discriminatory features were selected by recursive feature elimination and used to train a SVM classifier with a linear kernel. Compared to morphological, spatial, and volumetric features, tractography features achieved a high accuracy of 69.7% on the training cohort but a low accuracy of 35.7% on the validation cohort and 41.6% on the challenge cohort.

Islam et al. [76] employed PixelNet [77] for segmentation of brain tumor into three non-overlapping subregions followed by extraction of radiomic features including shape, volumetric, and first order features. A subset of 50 most predictive features was selected using cross validation and used to train an artificial neural network for prediction. The authors reported an accuracy of 46.8% on the challenge cohort.

Agravat et al. [1] were the winning team of the BraTS challenge in 2019. They used a 2D encoder-decoder architecture to segment brain tumors into three non-overlapping regions. A Random Forest regressor was trained with shape, volumetric and age features. Shape features were extracted only for the necrosis region. Volumetric features included the ratio of whole tumor volume to brain volume and the ratio of enhancing tumor, edema, and necrosis to whole tumor volume. The study reported an accuracy of 58.6% on the validation cohort and 57.9% on the challenge cohort.

Wang et al. [170] used an ensemble of 6 variants of U-Nets with loss function and postprocessing for segmentation of brain tumors into three non-overlapping subregions (PTE, NEC, and ENC). From the obtained segmentation maps, 13 shape and location, and 68 texture features were extracted. In addition to the radiomic features, the ratio of the second semi-axis of the tumor core to the second semi-axis of whole tumor was calculated and referred to as the relative invasiveness coefficient (RIC). Predictive

features were selected by recursive feature elimination with a random forest regressor. Three models were created to predict OS: (1) linear regressor with age only, (2) Random Forest with the 5 predictive radiomics features and age, and (3) Epsilon Support Vector Regressor with RIC and age. The authors reported an accuracy of 59.0% on the validation cohort and 56.0% on the challenge cohort with the third model.

Feng et al. [55] used an ensemble of six 3D U-Net architectures for segmentation of brain tumor into three non-overlapping subregions (PTE, NEC, and ENC). A linear regression model was trained with 6 volumetric features, extracted using multi-regional segmentation maps, and clinical features. The study reported an accuracy of 32.1% on the validation cohort and 61% on challenge cohort.

Pei et al. [130] proposed a 3D self-ensemble ResU-Net architecture for segmentation of brain tumor into three non-overlapping subregions (PTE, NEC, and ENC). 34 shape-features were extracted from the obtained multi-regional segmentation maps and ranked based on the feature importance attribute of a Random Forest classifier. The most predictive features were used to train a Random Forest regressor. The authors reported an accuracy of 55.2% on validation cohort and 43% on challenge cohort.

McKinley et al. [112] utilized a 3D-to-2D FCNN for overlapping segmentation of brain tumor volume i.e., Whole Tumor (WT), Tumor Core (TC), and active tumor (EC). Three features – number of distinct tumor components, number of tumor cores, and age – were used to train a fusion of linear regression and Random Forest classifiers. The study reported an accuracy of 61.7% on the challenge cohort.

Bommineni [22] used an ensemble of four 3D U-Nets, called Piece-Net, for non-overlapping segmentation of brain tumor volume. Radiomic and clinical features including volume, surface area, spatial location, and age were used to train a linear regression model. The study reported an accuracy of 37.9% on the validation cohort and 58.9% on the challenge cohort.

Asenjo and Solís [111] used an ensemble of four U-Net networks (three 2D U-Nets and

one 3D U-Net) for segmentation of brain tumor into three non-overlapping regions. The obtained multi-regional segmentation maps were used to extract a diverse set of radiomic features including first-order, shape, texture, and spatial features. Three models were independently learned for the OS classification task: (1) a RUSboosted decision tree classifier was trained using a subset of 24 predictive features obtained with chi-square test, (2) an SVM classifier with a quadratic kernel was trained using a subset of 10 predictive features obtained with MRMR method, and (3) a regression tree was trained using a subset of 29 predictive features obtained with F-test. Discrete label predictions were replaced with (continuous) survival days as follows: 150 days for short-term survivors, 376 days for medium-term survivors, and 796 days for long-term survivors. The final prediction was obtained by taking mean of the continuous values (survival days) of the three trained models. The study reported an accuracy of 61.7% on the challenge cohort.

Numerous studies have shown that, in comparison to classification models trained with handcrafted (radiomic) features, DL models reported poor predictive performance on BraTS validation and challenge cohorts [155, 66, 152, 3]. For instance, Akbar et al. [3] extracted deep features from 2D multi-parametric MRI scans by employing the modified versions of MobileNet V1 [69] and MobileNet V2 [146] architectures. Deep features, augmented with a clinical feature (age in years), were subsequently fed to a deep learning prediction module called survival prediction model (SPM). The study reported an accuracy of 31% on the validation cohort and 40.2% on the challenge cohort.

Two recent studies demonstrated instead strong performance of deep models for the OS classification task. Zhao et al. [185] used a deep learning framework, called Segmentation then Prediction (STP), based on 3D U-Net. The STP framework is composed of a segmentation module, which segments the brain tumor volume into overlapping sub-regions (i.e., WT, TC, EC), a local branch, which extracts features from whole tumor only, and a global branch which extracts features from the last layer of the segmenta-

tion module. Features from global and local branches are fused together to generate survival prediction. The study reported an accuracy of 65.5% on the validation cohort and 44.9% on the challenge cohort. Carmo et al. [25] employed a 3D U-Net with self-attention blocks for segmentation of brain tumor volume into overlapping subregions followed by prediction of OS class. The study reported an accuracy of 55.2% on the validation cohort and 46.7% on challenge cohort. It is important to note that the generalizability of deep models varied significantly between validation and challenge cohorts.

6.8 Our research goals

Radiomics-based approaches to OS prediction are impacted by choices made at each step of the radiomics process. These include the image acquisition and pre-processing parameters, the algorithms used for segmentation of tumor regions, the methods used to extract features and perform the analysis.

Our focus in this work is on the critical segmentation sub-process and on the evaluation of feature robustness. Radiomic features that are extracted from 3D Multiparametric MRI and that are used to predict OS are sensitive to the variability in tumor subregions segmentation algorithms. While many algorithms have been proposed for automatic segmentation of brain tumor sub-regions, no evidence is available about which algorithm is more appropriate in terms of radiomic performance. Additionally, all studies of feature robustness evaluated robustness with respect to manual or semiautomatic segmentation but did not assess the ultimate utility of these features for outcome prediction. One of our research goals is thus to identify MRI-based radiomic features that are robust to fully automated tumor segmentation and can be used to predict overall survival in HGGs.

In particular, we used the standard BraTS dataset of MR images provided by the Brain Tumor Segmentation Challenge (BraTS) [115, 11, 10, 12] and the shape, volumet-

ric, and spatial features commonly used by BraTS winning teams to reach the following research goals:

- Quantitatively evaluate the impact of state-of-the-art DL segmentation algorithms on radiomics-based prediction of OS in HGGs.
- Quantitatively evaluate the efficacy of multi-region segmentation maps, obtained using the STAPLE label fusion method [142], on radiomics-based prediction of OS in HGGs.
- Explore the efficacy of *6-subregions* and *21-subregions* radiomic models obtained using an anatomy-guided multi-regional segmentation of the brain tumor volume for the OS classification task.
- Provide a failure analysis of the considered multi-regional radiomic models for the OS classification task.
- Evaluate the robustness in terms of stability of radiomic features extracted from state-of-the-art DL algorithms.
- Compare the performance of stable and discriminatory features with the performance of discriminatory features alone for the OS classification task.

In the next chapter, we will describe the complete radiomics process we followed to evaluate the impact of selected state-of-the-art segmentation algorithms on radiomics-based prediction of OS in HGGs.

7

Evaluation of the impact of segmentation algorithms on OS prediction with multiregional radiomics

As discussed in the previous chapter, tumor subregion segmentation is a fundamental stage in the radiomics process. While many algorithms have been proposed for automatic segmentation of brain tumor sub-regions, there has been no significant analysis of their effect on radiomic performance. OS prediction is notoriously sensitive to the variability in tumor subregions segmentation algorithms so exploring their impact is an important goal.

In this chapter, we present the experimental methodology we followed to evaluate the robustness of OS prediction to variations in the automatic segmentation of brain

tumor volume in radiomics analysis. We compared five state-of-the-art Deep Learning (DL) algorithms (Dong 2D U-Net, Wang 2.5D CNN, Isensee 3D U-Net, HDC-Net, and E_1D_3 3D U-Net) as well as the STAPLE label fusion method [142], which is used to fuse the segmentation labels obtained from the five DL segmentation algorithms. All the implementation work was done in Python 3.6 using the following open-source packages: scikit-learn [127], N4ITK bias field correction [164], ANTs [7], PyRadiomics¹ [165], Pandas [113], Nibabel², and STAPLE fusion³ [142].

7.1 Experimental methodology

7.1.1 Data

As input to the radiomics process, we made use of the publicly available BraTS 2020 dataset of 3D multiparametric MRI scans [115, 11, 10, 12]. The training cohort consists of 369 subjects with preoperative 3D multiparametric MRI scans (including T1, T2, T1ce, and FLAIR sequences). Manual segmentation of tumor subregions (including peritumoral edema, non-enhancing core, and enhancing core) is included and confirmed by expert neuroradiologists [115]. Out of 369 subjects, 76 are low-grade gliomas (LGGs) cases and 293 are high-grade gliomas (HGGs) cases. Out of 293 HGGs, complete survival information was provided for 236 subjects and Gross Tumor Resection (GTR) status was provided only for 118 subjects. Of the 118 subjects, 42 are classified as short-term survivors, 30 are medium-term survivors, and 46 are long-term survivors. The validation cohort consists of 125 subjects and GTR status is provided for 29 subjects only. Unlike the training cohort, the validation cohort only contained preoperative 3D multiparametric MRI scans (including T1, T2, T1ce, and FLAIR sequences), and did not include manual segmentation of tumor subregions or survival information. Predictions

¹<https://pyradiomics.readthedocs.io>

²<https://github.com/nipy/nibabel>

³<https://github.com/FETS-AI/LabelFusion>

on the validation cohort can only be evaluated online on the CBICA⁴ portal. The challenge cohort consists of 166 subjects and is not publicly available for experiments and evaluation. The BraTS 2020 dataset also includes subjects from the The Cancer Imaging Archive (TCIA) [35, 148] and provides a name mapping file that matches the BraTS 2020 subject IDs with the TCIA subject IDs. With the help of matched TCIA subject IDs, we managed to extract survival information and clinical variables of an additional 31 HGGs from the validation cohort of the BraTS 2020 dataset. Of the 31 subjects, 16 are short-term survivors, 3 are medium-term survivors, and 12 are long-term survivors.

To summarize, we used the 3 following data cohorts in our work:

- Training cohort (118 subjects)
- Testing cohort A (31 subjects)
- Testing cohort B (29 subjects)

Manual segmentation of tumor subregions is only available for the training cohort. Survival information is only available for the training cohort and testing cohort A. Table 7.1 summarizes demographic and clinical characteristics of the training and testing cohorts.

7.1.2 Preprocessing

The 3D MRI scans for each subject were already skull-stripped, registered to T1ce scan, and resampled to an isotropic $1 \times 1 \times 1 \text{ mm}^3$ resolution [11]. The 3D MRI T1 scan for each subject was preprocessed using the N4ITK bias field correction algorithm [164], which is a recommended pre-processing step before performing any medical image processing task such as image registration [118].

⁴CBICA Image Processing Portal: <https://ipp.cbica.upenn.edu/>

Table 7.1: Overview of the training and testing cohorts (A and B) used in the overall survival classification task

| Characteristics | Training Cohort | Testing Cohort A | Testing Cohort B |
|--|-----------------|------------------|------------------|
| Patient demographic | | | |
| No. of patients | 118 | 31 | 29 |
| Patient distribution | | | |
| CBICA UPenn | 94 | - | 15 |
| TCIA | 17 | 31 | - |
| Others ¹ (NA, MDA, UAB, WashU) | 7 | - | 14 |
| Imaging data | | | |
| 3D multiparametric MRI scans (T1, T1ce, T2, and FLAIR) | ✓ | ✓ | ✓ |
| Ground truth Segmentation masks | ✓ | ✗ | ✗ |
| Clinical Information | | | |
| Age (years) ($p = 0.252$) ² ($p = 0.115$) ³ | | | |
| Range | 27.8-86.6 | 17.0 - 80.0 | 21.7 - 85.6 |
| Mean | 61.9 | 58.4 | 57.3 |
| Median | 63.5 | 58 | 58 |
| 1 Standard deviation | 12.0 | 15.5 | 14.3 |
| Survival groups ($p = 0.40$) ⁴ | | | |
| Range (days) | 12-1767 | 16-1215 | - |
| Mean (days) | 446.4 | 390.8 | - |
| Median (days) | 374.5 | 293.7 | - |
| 1 Standard deviation (days) | 343.8 | 314.4 | - |
| Short-term [<10 days] | 42 | 16 | - |
| Medium-term [10 - 15 months] | 30 | 3 | - |
| Long-term [>15 months] | 46 | 12 | - |
| Notes: | | | |
| ¹ Information is not shared by the BraTS 2020 organizers | | | |
| ² p-value for statistical comparison of age between training cohort and testing cohort A | | | |
| ³ p-value for statistical comparison of age between training cohort and testing cohort B | | | |
| ⁴ p-value for statistical comparison of survival between training cohort and testing cohort A | | | |

7.1.3 Brain tumor segmentation

Manual segmentations of tumor subregions are already provided for the training cohort by BraTS challenge organizers. For testing cohorts (A and B), we automatically generated the segmentation of the brain tumor volume using five state-of-the-art CNNs discussed in detail in the previous chapter (Dong 2D U-Net, Wang 2.5D CNN, Isensee 3D U-Net, HDC-Net, and E_1D_3 3D U-Net) after training them on the BraTS 2020 training data.

We also employed the STAPLE fusion method [142] to fuse the segmentation labels

obtained from the CNNs algorithms.

The segmentations were performed on a system with 64 GB RAM, and an NVIDIA RTX 2080Ti 11 GB GPU using the Tensorflow framework. Configuration and hyperparameters for the five segmentation architectures are presented in Table 7.2.

Table 7.2: Configuration and hyperparameters of the five CNNs used for automatic segmentation of brain tumor volume (data provided by Syed. Talha Bukhari).

| Network | Dong 2D U-Net | Wang 2.5D CNN | Isensee 3D U-Net | HDC-Net | ED3DU-Net |
|---|---|--|--|---|---|
| Architecture | 2D U-Net | Three 2.5D Anisotropic CNNs (W-Net, T-Net, and E-Net) in cascade | 3D U-Net with Deep supervision | 2.5D U-Net | 3D U-Net |
| Activation | ReLU | P-ReLU | Leaky-ReLU (0.01) | ReLU | Leaky-ReLU (0.01) |
| Batch size | 10 | 5 (Same for three CNNs in cascade) | 2 | 8 | 2 |
| Initialization | He-normal | Truncated Normal | He-normal | He-normal | He-normal |
| Input size/ Output size | $240^2/240^2$ | W-Net: $19 \times 144^2/11 \times 144^2$ T-Net: $19 \times 64^2/11 \times 64^2$ E-Net: $19 \times 64^2/11 \times 64^2$ | $128^3/128^3$ | $128^3/128^3$ | $96^3/96^3$ |
| Learning Rate policy ¹ | Polynomial decay (batch-wise) $\eta_0 = 10^{-4}$ $\eta_{end} = 10^{-7}$ $\gamma = 1.2$ | Constant (10^{-3}) | Polynomial decay (epoch-wise) $\eta_0 = 0.01$ $\gamma = 0.9$ | Polynomial decay (epoch-wise) $\eta_0 = 10^{-3}$ $\gamma = 0.9$ | Polynomial decay (epoch-wise) $\eta_0 = 10^{-2}$ $\gamma = 0.9$ |
| Optimizer | Adam | Adam | SGD + Nesterov (0.99) | Adam (AMSGrad variant) | SGD+Nesterov (0.99) |
| Loss | Soft Dice | Soft Dice | Soft Dice + Cross Entropy | Generalized Soft Dice | Soft Dice + Cross Entropy |
| Regularization | - | $L_2(10^{-7})$ | $L_2(3 \times 10^{-5})$ | $L_2(10^{-5})$ | $L_2(10^{-6})$ |
| Total Training iterations (Gradient-decent updates) | 50k (100 epochs) | 20k (per network) | 250k (1000 epochs) | 37.35k (900 epochs) | 125k (500 epochs) |
| # Parameters | 34.5 million | W-Net: 0.21 million T-Net: 0.21 million E-Net: 0.20 million | 31.2 million | 0.29 million | 34.9 million |
| Training Time ² | ~110 hours | W-Net (single-view): ~84 hours T-Net (single-view): ~84 hours E-Net (single-view): ~20 hours | ~101 hours | ~110 hours | ~48 hours |
| Test-time Augmentation | ✓ | ✗ | ✓ | ✓ | ✓ |
| Morphological Post-processing | Morphological closing, cluster thresholding | ✗ | ✗ | ✗ | ✓ |

Notes:
¹ For definition of variables consult Table 1 in [24].
² Please note that training time also depends on the GPU system used for training. HDC-Net was trained on a dual-GPU system whereas remaining CNNs were trained on a single-GPU system.

7.1.4 Tumor subregion segmentation models

We used the results of the segmentation algorithms to build 4 different tumor subregion segmentation models that could be used in the successive feature extraction stage of the radiomics process. The four resulting segmentation models are called *Whole Tumor* (WT) model, *3-subregions* model, *6-subregions* model, and *21-subregions* model.

The four segmentation models belong to two main categories of brain tumor segmentation models: (1) physiology-based models and (2) anatomy-based models.

Physiology-based segmentation model

In physiology-based models, the brain tumor is divided into three non-overlapping subregions (PTE, NEC, and ENC), which are obtained directly as the results of the segmentation algorithms. These tumor subregions can potentially provide better features that are consistent with the prognosis of the tumor. In this work, the physiology-guided segmentation model based on the use of the three separate non-overlapping subregions is referred as a *3-subregions* model. In addition, we defined the *Whole Tumor* (WT) model as the combination of the three non-overlapping subregions, which we also used in the process needed to create the anatomy-based segmentation models.

Anatomy-based segmentation models

In anatomy-based models, the brain tumor is subdivided into anatomical regions with the help of a pre-defined Harvard-Oxford subcortical atlas with 21 labeled anatomical regions [43]. Anatomy-based segmentation is obtained in four steps:

1. The Harvard-Oxford subcortical atlas is registered into subject space using diffeomorphic registration. To do this we used the SyNOnly algorithm as implemented in the ANTs (Advanced Normalization Tools) package [7]. SyNOnly was initialized with the output of affine registration and used mutual information as a cost

function.

2. The Whole Tumor (WT) mask is overlaid with the registered atlas to extract the tumor-affected anatomical regions.
3. Volumes of tumor-affected anatomical regions are computed and then ranked in descending order.
4. Finally, the top-K anatomical subregions that combine to occupy more than 85% of WT volume are retained.

In this work, we refer to the resulting segmentation model as *6-subregions* segmentation model where $6 (= K)$ is the number of subregions selected in step 4. For comparison, we also used the *21-subregions* segmentation model obtained in step 2.

Given a segmentation model, i.e., *WT*, *3-subregions*, *6-subregions*, or *21-subregions*, one can extract region-specific radiomic features for classification. Figure 7.1 shows the two segmentation models: physiology based segmentation, and anatomy based segmentation with 6 subregions.

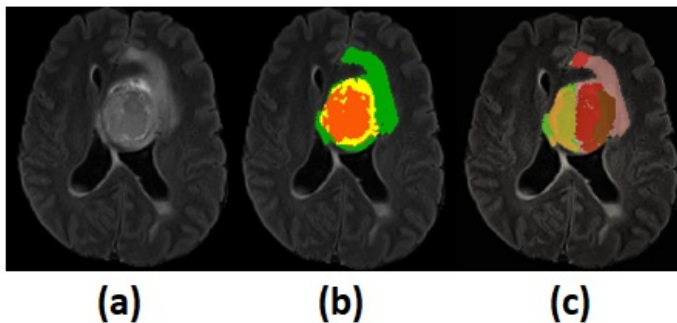


Figure 7.1: (a) axial slice of FLAIR scan overlain with (b) physiology-based manual segmentation including **peritumoral edema** (green), **enhancing tumor** (yellow), and **non-enhancing tumor** (orange), and (c) anatomy-based tumor segmentation with 6 -subregions model including **left cerebral cortex**, **right cerebral cortex**, **left lateral ventricle**, **right lateral ventricle**, **left cerebral white matter**, and **right cerebral white matter**. Note: This figure should be visualized in color.

7.1.5 Radiomic feature extraction

To compare the power of the different segmentation algorithms and models, we extracted radiomic features, using the PyRadiomics software package [165], from WT only (WT radiomics model), from the three non-overlapping subregions PTE, NEC, and ENC (3-subregions radiomics model), from the left and right cerebral cortex, left and right cerebral white matter, and left and right lateral ventricle (6-subregions radiomics model), and from 21 anatomical subregions provided by the registered Harvard-Oxford subcortical atlas (21-subregions radiomics model). We extracted the following set of (radiomic) features:

- **Shape features** include volume and surface area of each subregion. For instance, in the 3-subregions radiomics model, we extracted volume and surface area of peritumoral edema, non-enhancing core and enhancing core. In the OS classification task, shape features have been shown to provide insight in tumor behavior [136, 2, 22, 4, 130, 124]. Several studies report that tumor volume and surface area are strong predictors of survival in patients with glioblastoma [115, 170, 22, 55]. A large tumor volume reflects severity of tumor and is associated to poor prognosis and shorter survival times [115, 55]. We extracted 2 shape features for the WT radiomics model, 6 shape features for the 3-subregions radiomics model, 12 shape features for the 6-subregions radiomics model, and 42 shape features for the 21-subregions radiomics model.
- **Spatial features** capture the location of the tumor within the brain. More specifically, we extracted (a) coordinates (in 3D) of the centroid of the WT with respect to the brain mask and (b) the Euclidean distance between the centroid of the WT and the centroid of the brain mask. Brain mask is defined as the non-zero region in the 3D FLAIR sequence. Spatial features have been shown to be predictive for survival prediction task [27, 136, 22].

- **Demographic features** consist of the age (in years) of the subject, which is provided in the BraTS 2020 dataset.

Overall, a total of 7 features were obtained for the *WT* radiomics model, 11 features for the *3-subregions* radiomics model, 17 features for the *6-subregions* radiomics model, and 47 features for the *21-subregions* radiomics model. A summary of the considered radiomic features is provided in Table 7.3.

For the training cohort, radiomic features were extracted from 3D mpMRI scans using manual segmentations provided with the BraTS 2020 dataset. For testing cohorts (A and B), radiomic features were obtained using the six segmentation networks (five CNNs and one STAPLE-fused segmentation) presented in Section 6.5.

Every feature vector from the training cohort was independently normalized (i.e., transformed to z-scores) by subtracting the mean and dividing by the standard deviation. Features from testing cohorts A and B were normalized using the mean and standard deviation of the training cohort. No feature selection was performed for the 4 considered radiomic models, i.e. all features were used in the successive stage of the radiomics process.

Table 7.3: Summary of radiomic features extracted for four radiomic models, namely, *WT* radiomics model, *3-subregions* radiomics model, *6-subregions* radiomics model, and *21-subregions* radiomics model

| Feature Types | Feature Names | No of Features |
|---|--|------------------------------------|
| Clinical features | Age | 1 |
| Spatial features | Centroid of the WT, (Euclidean) Distance between the (centroid of) WT and the (centroid of) the brain | 4 |
| Shape features (WT radiomics model) | Volume and Surface Area of Whole Tumor | 2 |
| Shape features (3-subregions radiomics model) | Volume and Surface Area of Peritumoral Edema (PTE) Enhancing Core (ENC) and Non-Enhancing Core (NEC) | 6 (2 features * 3 subregions) |
| Shape features (6-subregions radiomics model) | Volume and Surface Area of Right Cerebral Cortex (RCC), Left Cerebral Cortex (LCC), Left Lateral Ventricle (LLV), Right Lateral Ventricle (RLV), Left Cerebral White Matter (LCWM) , Right Cerebral White Matter (RCWM) | 12 (2 features * 6 subregions) |
| Shape features (21-subregions radiomics model) | Volume and Surface Area of 21 Subcortical Regions defined by a registered Harvard-Oxford subcortical atlas Regions | 42 (2 features * 21 subregions) |

7.1.6 OS prediction: model training and inference

We used a Random Forest (RF) classifier to predict the survival class of patients based on the extracted radiomic features. In general, a RF is an ensemble of decision trees that can be used for different tasks such as classification, regression, and others [100]. Ensemble models like RF are preferable to individual models because they provide higher accuracy and better generalizability. Using a RF for classification tasks in medical image analysis is a good choice because it is an effective method in handling multi-class problems. Indeed, there is evidence in the literature that RF is the most effective and stable method for predicting overall survival in glioma patients [125, 174].

For the training phase, random forest classifiers ($N = 100$) were trained on the training cohort comprising of 118 subjects with GTR status. Hyperparameters of each random forest classifier were set as follows: (*no_of_estimators* = 200, *max_features* = *auto*, *class_weight* = *balanced*, *criterion* = *gini*).

For the inference phase, a soft voting method was adopted to unify the outputs of N random forest classifiers (with uniform weighting scheme) and generate a single prediction of OS class for each subject.

7.1.7 Evaluation

Metrics for segmentation performance

Performance of the six segmentation algorithms (the five CNNs and the STAPLE-fusion method) was quantified using Dice Similarity Coefficient (DSC) [44] and Hausdorff distance metric (HD-95) [73]. DSC quantifies the overlap between predicted and reference segmentation maps while HD-95 measures the degree of mismatch between the predicted and reference segmentation maps.

Consider a predicted binary segmentation map $x = [x_1, \dots, x_N]^T$ and a binary reference segmentation map $y = [y_1, \dots, y_N]^T$, with $(x_i, y_i) = 0, 1$. The DSC and

HD-95 metrics for $i = 1, \dots, N$, are computed as

$$DSC(x, y) = \frac{2|x \cap y|}{|x| + |y|} = \frac{2 \sum_{i=1}^N x_i y_i}{\sum_{i=1}^N x_i + \sum_{i=1}^N y_i} \quad (7.1)$$

$$HD(x, y) = \max \left(\max_{i \in x} \left(\min_{j \in y} d(i, j) \right), \max_{j \in y} \left(\min_{i \in x} d(i, j) \right) \right)$$

where $DSC \in [0, 1]$ and $HD \in [0, \infty)$.

The six segmentation algorithms were ranked based on the Final Ranking Score (FRS) and statistical significance (of ranking) was calculated using a random permutation test [24]. A lower FRS value means a higher ranking. FRS is calculated in three steps:

1. Rank each subject segmentation with respect to others by using the HD and DSC of three subregions (WT, TC and EC) in the six segmentation networks. This results in 36 different rankings (3 subregions x 2 metrics x 6 networks).
2. Take the mean of the 6 rankings of each segmentation network obtaining a cumulative rank for the 6 segmentation networks for each subject.
3. Compute the FRS via averaging of cumulative ranks of all subjects.

Metrics for radiomics performance

On testing cohort A (31 subjects), predictive performance of radiomic models was quantified using area under the receiver operating curve (AUC) and area under the precision-recall curve (AUPRC). On testing cohort B (29 subjects), predictive performance of radiomic models could only be quantified with the accuracy (acc) metric on the CBICA online portal. Accuracy measures the correctly classified and misclassified cases for each class:

$$Accuracy = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \quad (7.2)$$

Classification performance in case of an imbalanced dataset is measured by the Area Under the Curve (AUC) and Area Under the Precision Recall Curve (AUPRC). AUC plots the True Positives Rate (TPR) vs False Positive Rate (FPR). TPR is defined as $TPR = \frac{t_p}{t_p+f_n}$ while FPR is defined as $FPR = \frac{f_p}{f_p+t_n}$.

AUPRC is the curve between precision and recall, with precision defined as $precision = \frac{t_p}{t_p+f_p}$ and recall defined as $recall = \frac{t_p}{t_p+f_n}$.

Stability of the radiomic models was quantified with relative standard deviation (RSD) calculated as a ratio of standard deviation to the mean of AUC. A lower value of RSD corresponds to higher stability of the radiomic models.

$$RSD = \frac{\sigma_{AUC}}{\mu_{AUC}} * 100 \quad (7.3)$$

Statistical analysis of demographic data (in Table 7.1) was performed using the student t-test. A $p < 0.05$ was considered statistically significant and a $p < 0.001$ was considered statistically highly significant.

7.2 Results

7.2.1 Clinical characteristics

Table 7.1 displays clinical characteristics of the training cohort and testing cohorts. The median age of the training cohort, testing cohort A, and testing cohort B were 63.5, 58, and 58 years respectively. No statistical difference was found in age between the training cohort and testing cohort A ($p = 0.252$) and the training cohort and testing cohort B ($p = 0.115$). The median overall survival (in days) for the training cohort and testing cohort A were 375 days and 294 days respectively. While the training cohort was balanced across three survival groups, i.e., short-term (42 subjects), medium-term (30 subjects), and long-term (46 subjects) survivors, testing cohort A had a sparse presence

of medium-term survivors – only 3 subjects out of 31. No statistical difference was found in overall survival days between the training cohort and testing cohort A ($p = 0.40$). Survival information was not made publicly available for testing cohort B by the BraTS 2020 organizers.

7.2.2 Segmentation algorithm performance

Performance of the six considered segmentation algorithms, for testing cohorts A and B combined (60 subjects), is summarized in Table 7.4. We used Final Ranking Score (FRS) to unify the 6 segmentation performance metrics (i.e., DSC and HD-95 scores for three subregions each) for each subject in testing cohorts A and B.

In terms of FRS, Isensee 3D U-Net was ranked significantly higher ($p < 0.001$) in comparison to the remaining CNNs for brain tumor segmentation. Isensee 3D U-Net obtained the highest DSC scores for WT ($DSC = 91.5$), TC ($DSC = 90.9$), and EN ($DSC = 87.0$) subregions which quantifies overlap with manual segmentation maps. In terms of the HD-95 metric, Isensee 3D U-Net was quite close in performance to HDC-Net ($\Delta HD_{avg} = 0.07$) and much better than E1D3 3D U-Net ($\Delta HD_{avg} = 1.2$), Wang 2.5D CNN ($\Delta HD_{avg} = 1.43$), and Dong 2D U-Net ($\Delta HD_{avg} = 1.53$).

The STAPLE-fusion method ranked second, in terms of FRS, but not significantly lower than Isensee 3D U-Net ($p = 0.205$). However, the STAPLE-fusion method was ranked significantly higher than Dong 2D U-Net ($p < 0.001$), Wang 2.5D CNN ($p < 0.001$), HDC-Net ($p < 0.001$), and E1D3 3D U-Net ($p < 0.001$). Compared to the five CNNs (individually), the STAPLE-fusion method reported the lowest HD-95 scores which measure the degree of mismatch between manual and predicted segmentation maps. Figure 7.1 shows the predicted multi-class segmentation maps obtained with six segmentation networks for three subjects, one from each survival class, in testing cohort A.

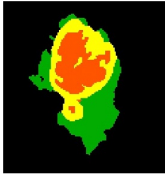
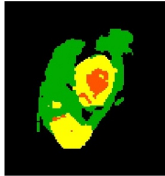
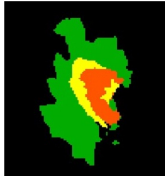
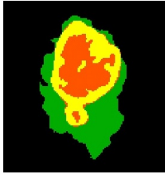
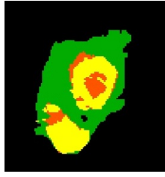
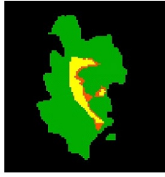
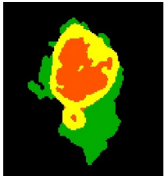
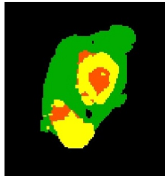
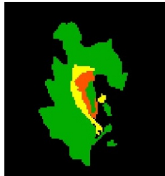
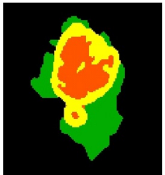
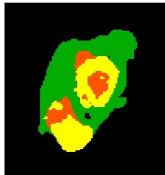
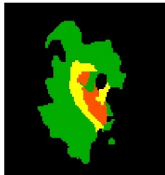
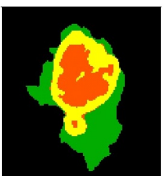
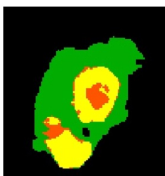
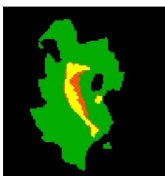
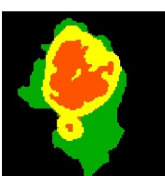
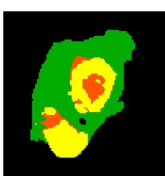
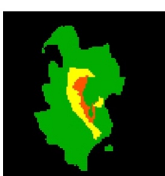
| Segmentation network | Overall Survival group | | |
|----------------------|---|---|---|
| | Short-term survivor | Medium-term survivor | Long-term survivor |
| Dong 2D U-Net |  |  |  |
| Wang 2.5D CNN |  |  |  |
| Isensee 3D U-Net |  |  |  |
| HDC-Net |  |  |  |
| EiDs 3D U-Net |  |  |  |
| STAPLE Fusion |  |  |  |

Figure 7.1: Automatically segmented tumor subregions from the five CNNs-based segmentation networks and the STAPLE fusion method. Label legend: Peritumoral Edema (green), Enhancing Core (yellow), Non-enhancing Core (orange).

Table 7.4: Performance of the five CNNs-based segmentation networks and the STAPLE-fusion method on testing cohorts A and B (60 subjects). **Bold font** indicates best scores for overlapping subregions (WT, TC, and EC)

| Segmentation Network | Dice Similarity Coefficient (%) | | | Hausdorff Distance | | | Final Ranking Score (FRS) |
|--|---------------------------------|-------------------|-------------------|--------------------|------------------|------------------|---------------------------|
| | WT | TC | EC | WT | TC | EC | |
| Dong 2D U-Net | 90.4 ± 6.5 | 87.3 ± 9.9 | 84.1 ± 9.4 | 5.8 ± 9.0 | 6.5 ± 8.6 | 3.2 ± 5.5 | 6** |
| Wang 2.5D CNN | 90.6 ± 5.5 | 89.3 ± 8.7 | 85.2 ± 10.0 | 6.6 ± 10.0 | 5.7 ± 8.7 | 2.9 ± 4.8 | 5** |
| Isensee 3D U-Net | 91.5 ± 5.5 | 90.9 ± 6.7 | 87.0 ± 7.3 | 4.4 ± 5.6 | 4.4 ± 8.5 | 2.1 ± 1.9 | 1 |
| HDC-Net | 90.8 ± 5.4 | 90.1 ± 7.3 | 85.9 ± 8.4 | 4.3 ± 4.4 | 4.5 ± 8.0 | 2.1 ± 1.3 | 3** |
| E ₁ D ₃ 3D U-Net | 91.4 ± 4.9 | 89.7 ± 9.0 | 85.9 ± 9.1 | 5.5 ± 7.8 | 5.6 ± 10.0 | 3.4 ± 6.3 | 4** |
| STAPLE Fusion | 91.4 ± 4.8 | 90.6 ± 7.6 | 86.7 ± 7.7 | 4.1 ± 3.4 | 4.4 ± 8.1 | 2.0 ± 1.3 | 2 |

7.2.3 Radiomics models evaluation on Testing Cohort A (31 subjects)

Performance measures (AUC, AUPRC, and RSD) for the four considered radiomic models are summarized in Table 7.5.

Table 7.5: Performance of the 4 radiomic models on testing cohort A (31 subjects). **Bold font** indicates best performance achieved for each radiomic model.

| Segmentation Network | Performance Metric | WT radiomic model | 3-subregions radiomic model | 6-subregions radiomic model | 21-subregions radiomic model |
|--|--------------------|----------------------------|-----------------------------|-----------------------------|------------------------------|
| Dong 2D U-Net | AUC | 0.70 (0.71, 0.66, 0.69) | 0.75 (0.75, 0.46, 0.77) | 0.71 (0.77, 0.48, 0.70) | 0.70 (0.68, 0.43, 0.78) |
| | AUPRC | 0.58 | 0.66 | 0.51 | 0.57 |
| Wang 2.5D CNN | AUC | 0.68 (0.68, 0.44, 0.7) | 0.75 (0.66, 0.48, 0.87) | 0.70 (0.71, 0.42, 0.74) | 0.70 (0.65, 0.43, 0.78) |
| | AUPRC | 0.53 | 0.68 | 0.51 | 0.56 |
| Isensee 3D U-Net | AUC | 0.70 (0.72, 0.38, 0.72) | 0.71 (0.7, 0.31, 0.75) | 0.73 (0.75, 0.44, 0.78) | 0.72 (0.69, 0.45, 0.72) |
| | AUPRC | 0.57 | 0.62 | 0.56 | 0.61 |
| HDC-Net | AUC | 0.69 (0.68, 0.45, 0.72) | 0.73 (0.67, 0.53, 0.82) | 0.71 (0.71, 0.45, 0.75) | 0.71 (0.67, 0.45, 0.78) |
| | AUPRC | 0.54 | 0.61 | 0.53 | 0.58 |
| E ₁ D ₃ 3D U-Net | AUC | 0.67 (0.68, 0.36, 0.69) | 0.72 (0.71, 0.35, 0.8) | 0.71 (0.73, 0.42, 0.76) | 0.72 (0.69, 0.51, 0.79) |
| | AUPRC | 0.54 | 0.64 | 0.57 | 0.60 |
| STAPLE Fusion | AUC | 0.68 (0.69, 0.42, 0.71) | 0.74 (0.7, 0.45, 0.82) | 0.70 (0.75, 0.40, 0.74) | 0.71 (0.69, 0.43, 0.77) |
| | AUPRC | 0.55 | 0.64 | 0.51 | 0.59 |

Note: The micro-AUC of the three classes is displayed as an ordered triplet, (short-term survivor, medium-term survivor, and long-term survivor) below the weighted average AUC value.

WT radiomics model

For the WT model, results showed that Dong 2D U-Net and Isensee 3D U-Net reported highest predictive performance (AUC = 0.70 and AUPRC = 0.58) and E_1D_3 3D U-Net showed lowest predictive performance (AUC = 0.67 and AUPRC = 0.54). While Isensee 3D U-Net showed strong predictive power for short-term survivors (AUC = 0.72) and long-term survivors (AUC = 0.72), its performance dropped considerably for medium-term survivors (AUC = 0.38). Dong 2D U-Net displayed best predictive performance for medium-term survivors (AUC = 0.66) while maintaining high predictive performance on short-term survivors (AUC = 0.71) and long-term survivors (AUC = 0.69). The stability of the WT radiomics model was 1.52 as measured with RSD, across the six segmentation methods. The STAPLE-fusion method marginally exceeded the predictive performance of E_1D_3 3D U-Net and was inferior to the remaining segmentation networks.

3-subregions radiomics model

For the 3-subregions model, results showed that Wang 2.5D CNN and Dong 2D U-Net reported highest predictive performance (AUC = 0.75 and AUPRC = 0.68) and Isensee 3D U-Net showed lowest predictive performance (AUC = 0.71 and AUPRC = 0.62). While Dong 2D U-Net showed strong predictive power for short-term survivors (AUC = 0.75) and long-term survivors (AUC = 0.77), its performance dropped considerably for medium-term survivors (AUC = 0.46). HDC-Net displayed best predictive performance for medium-term survivors (AUC = 0.53) while maintaining high predictive performance on long-term survivors (AUC = 0.82) and short-term survivors (AUC = 0.67). The stability of the 3-subregions radiomics model was 1.99 as measured with RSD, across the six segmentation methods. STAPLE-fusion method exceeded the predictive performance of E_1D_3 3D U-Net, HDC-Net, and Isensee 3D U-Net and was inferior to the remaining segmentation networks.

6-subregions radiomics model

For the 6-subregions model, results showed that Isensee 3D U-Net reported highest predictive performance (AUC = 0.73 and AUPRC = 0.56) and Wang 2.5D CNN showed lowest predictive performance (AUC = 0.70 and AUPRC = 0.51). Dong 2D U-Net showed best predictive performance for medium-term survivors (AUC = 0.48) while maintaining strong performance on short-term survivors (AUC = 0.77) and long-term survivors (AUC = 0.70). The stability of the 6-subregions radiomics model, across the six segmentation methods, was 1.48. The predictive performance of STAPLE-fusion method was similar to Wang 2.5D U-Net and inferior to the remaining segmentation networks.

21-subregions radiomics model

For the 21-subregions model, results showed that Isensee 3D U-Net and E_1D_3 3D U-Net reported highest predictive performance (AUC = 0.72 and AUPRC = 0.61) and Dong 2D U-Net and Wang 2.5D CNN showed lowest predictive performance (AUC = 0.70 and AUPRC = 0.57). E1D3 3D U-Net showed best predictive performance for medium-term survivors (AUC = 0.51) while maintaining strong performance on short-term survivors (AUC = 0.69) and long-term survivors (AUC = 0.79). The stability of the 21-subregions radiomics model, across the six segmentation methods, was 1.39. STAPLE-fusion method marginally exceeded the predictive performance of Dong 2D U-Net and Wang 2.5D U-Net and was inferior to the remaining segmentation networks.

7.2.4 Failure analysis

Finally, we performed failure analysis on the radiomics models by studying subjects which were misclassified by a majority of segmentation schemes. More specifically, for each radiomics model, we identified subjects misclassified with (a) all six segmentation

schemes (0-6), (b) five segmentation schemes (1-5), and (c) four segmentation schemes (2-4).

The analysis for the WT radiomics model, 3-subregions radiomics model, and 6-subregions radiomics model revealed that 16 (distinct) subjects were misclassified for at least one radiomics model. Out of 16 subjects, 8 were short-term survivors, 3 were medium-term survivors, and 5 were long-term survivors. Figure 7.2(A) shows a Venn diagram which distributes the 16 misclassified subjects across three radiomic models. 8 out of 16 subjects were misclassified by all three radiomic models.

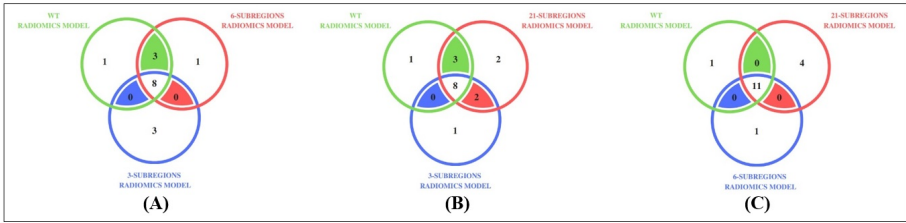


Figure 7.2: Distribution of misclassified subjects in (A) *WT* radiomics model, *3-subregions* radiomics model, and *6-subregions* radiomics model (B) *WT* radiomics model, *3-subregions* radiomics model, and *21-subregions* radiomics model (C) *WT* radiomics model, *6-subregions* radiomics model, and *21-subregions* radiomics model, on testing cohort A (31 subjects)

The analysis for the WT radiomics model, 3-subregions radiomics model, and 21-subregions radiomics model revealed that 17 (distinct) subjects were misclassified for at least one radiomics model. Out of 17 subjects, 9 were short-term survivors, 3 were medium-term survivors, and 5 were long-term survivors. Figure 7.2(B) shows a Venn diagram which distributes the 17 misclassified subjects across three radiomic models. 8 out of 17 subjects were misclassified by all three radiomic models.

The analysis for the 6-subregions radiomics model and 21-subregions radiomics model revealed that 16 (distinct) subjects were misclassified for at least one radiomics model. Out of 16 subjects, 9 were short-term survivors, 2 were medium-term survivors, and 5 were long-term survivors. Figure 7.2(C) shows a Venn diagram which distributes the misclassified subjects across WT radiomics model, 6-subregions radiomics model, and

21-subregions radiomics model. Most subjects (11 out of 12) misclassified by the WT radiomics model also failed with the 6-subregions radiomics model and 21-subregions radiomics model.

7.2.5 Radiomics models evaluation on Testing Cohort B (29 subjects)

Classification accuracy for the four radiomic models on Testing Cohort B is summarized in Table 7.6. Testing cohort B could only be evaluated online on the CBICA portal which only reports classification accuracy. In terms of accuracy, the four radiomics models reported superior (and matched) performance with multiple segmentation schemes. For the WT radiomics model, HDC-Net, E_1D_3 3D U-Net, and STAPLE-fusion obtained the highest accuracy (48.3%). For the 3-subregions radiomics model, the highest accuracy (44.8%) was obtained with E_1D_3 3D U-Net. For the 6-subregions radiomics model, the highest accuracy of 48.3% was obtained with Dong 2D U-Net, Isensee 3D U-Net, E_1D_3 3D U-Net, and STAPLE-fusion. For the 21-subregions radiomics model, the highest accuracy of 51.7% was obtained with Dong 2D U-Net, and E_1D_3 3D U-Net. Amongst the six segmentation schemes, E_1D_3 3D U-Net obtained the highest accuracy for the WT radiomics model (48.3%), the 3-subregions radiomics model (44.8%), the 6-subregions radiomics model (48.3%), and the 21-subregions radiomics model (51.7%).

Table 7.6: Classification accuracy of the four radiomics models on testing cohort B (29 subjects)

| Segmentation Network | Accuracy (%) | | | |
|----------------------|-------------------|-----------------------------|-----------------------------|------------------------------|
| | WT radiomic model | 3-subregions Radiomic Model | 6-subregions radiomic model | 21-subregions radiomic model |
| Dong 2D U-Net | 44.8 | 41.4 | 48.3 | 51.7 |
| Wang 2.5D CNN | 44.8 | 41.4 | 41.4 | 37.9 |
| Isensee 3D U-Net | 44.8 | 41.4 | 48.3 | 44.8 |
| HDC-Net | 48.3 | 37.9 | 44.8 | 44.8 |
| E_1D_3 3D U-Net | 48.3 | 44.8 | 48.3 | 51.7 |
| STAPLE Fusion | 48.3 | 41.4 | 48.3 | 41.4 |

7.3 Discussion

In this work, we explored the efficacy of four radiomic models – WT radiomics model, 3-subregions radiomics model, 6-subregions radiomics model, and 21-subregions radiomics model – for overall survival (OS) classification task in brain gliomas. The WT radiomics model extracts features from the WT region only. The 3-subregions radiomics model extracts features from three non-overlapping subregions of WT i.e., PTE, NEC, and ENC. The 6-subregions radiomics model extracts features from six anatomical regions overlapping with WT volume including left and right cerebral cortex, the left and right cerebral white matter, and the left and right lateral ventricle subregions. The 21-subregions radiomics model extracts features from 21 anatomical regions provided with Harvard-Oxford subcortical atlas (Table 7.7 for 21 anatomical regions). We also quantified the stability of radiomic models across six segmentation networks – five CNNs and one STAPLE-fusion method. The five CNNs include three 3D CNNs – Isensee 3D U-Net, E_1D_3 3D U-Net, and HDC-Net– one 2.5D CNN Wang 2.5D CNN, and one 2D CNN, Dong 2D U-Net. For each subject in testing cohorts A and B, the predicted segmentation maps from five CNNs were fused using the STAPLE-fusion method.

We benefitted from the publicly available BraTS 2020 and TCIA datasets and extracted three data cohorts – training cohort (118 subjects), testing cohort A (31 subjects), and testing cohort B (29 subjects). The training cohort comprised of HGGs with 3D multiparametric MRI scans and manual segmentation of brain tumor volume into three non-overlapping subregions i.e., PTE, NEC, and ENC. Testing cohorts A and B also comprised of HGGs but only included 3D multiparametric MRI scans. While the training cohort was reasonably balanced for the three survival classes – short-term survivors (42), medium-term survivors (30), and long-term survivors (46) – testing cohort A had sparse representation of medium-term survivors with only 3 subjects out of 31. Overall survival information for testing cohort B was not available offline.

Table 7.7: The 21 subregions defined by the Harvard-Oxford subcortical atlas

| Label | Anatomical Region |
|--------------|-----------------------------|
| 1 | Left Cerebral White Matter |
| 2 | Left Cerebral Cortex |
| 3 | Left Lateral Ventricle |
| 4 | Left Thalamus |
| 5 | Left Caudate |
| 6 | Left Putamen |
| 7 | Left Pallidum |
| 8 | Brainstem |
| 9 | Left Hippocampus |
| 10 | Left Amygdala |
| 11 | Left Accumbens |
| 12 | Right Cerebral White Matter |
| 13 | Right Cerebral Cortex |
| 14 | Right Lateral Ventricle |
| 15 | Right Thalamus |
| 16 | Right Caudate |
| 17 | Right Putamen |
| 18 | Right Pallidum |
| 19 | Right Hippocampus |
| 20 | Right Amygdala |
| 21 | Right Accumbens |

Segmentation of brain tumor volume is the penultimate step in any radiomics framework for brain gliomas. For each subject in testing cohorts A and B, the brain tumor volume was segmented into three non-overlapping regions (PTE, NEC, and ENC) using the aforementioned six segmentation networks. Our results showed that 3D CNNs, including Isensee 3D U-Net, HDC-Net, and E_1D_3 3D U-Net, provided superior segmentation of brain tumor subregions by utilizing 3D contextual information in volumetric scans. Amongst the five CNNs employed for brain tumor segmentation, E_1D_3 3D U-Net had a large memory footprint (35 million trainable parameters) and shortest training time (48 hours) and HDC-Net had the fewest trainable parameter (0.29 million trainable parameters) with long training time (110 hours). The STAPLE fusion method significantly outperformed four (of the five) CNNs ($p < 0.001$) except for Isensee 3D

U-Net which was ranked higher ($p = 0.205$). Moreover, the STAPLE fusion method reported the lowest HD-95 scores which has been observed previously with ensemble methods [56, 61, 119, 177]. Isensee 3D U-Net superior performance is attributed to the fact that the underlying 3D U-Net architecture was carefully optimized by empirically tuning network and training parameters on the BraTS dataset.

Automatic segmentation of brain tumors and their subregions is a challenging task for several reasons, including heterogeneity of tumor shape and appearance, unclear tumor boundaries, lack of high-quality imaging data, unbalanced tumor tissue, and the presence of artifacts [102, 31, 104]. The WT radiomics model, the 6-subregions radiomics model, and the 21-subregions radiomics model required accurate segmentation of WT volume which, in terms of Dice score, was performed quite similarly by the six segmentation networks ($DSC : 90.4 - 91.5\%$). However, in terms of Hausdorff distance – which measures the largest segmentation error – segmentation of WT volume had a large variability across six segmentation networks ($HD - 95 : 4.1 - 6.6mm$). The 3-subregions radiomics model required accurate delineation of additional subregions including Tumor Core (TC) and Active Tumor (EC). The segmentation of TC subregion varied substantially across the six segmentation networks, in terms of DSC (87.3–90.9%) and HD-95 (4.4 – 6.5mm) metrics. The segmentation of EC subregion is increasingly difficult because of poor contrast, fragmented (physiologic) structure and low-contrast MR images as reflected by uncertainty information associated with the segmentation results [170]. This was exhibited by reduced segmentation accuracy ($DSC : 84.1-87.0\%$) across the six segmentation networks.

The four radiomic models were obtained by training Random Forest classifiers ($N = 100$ for each radiomics model) using shape, volumetric, spatial, and demographic features. Our results showed that the 3-subregions radiomics model reported superior predictive performance ($mean AUC = 0.73$), across the six segmentation networks, compared to the WT radiomics model ($mean AUC = 0.69$), the 6-subregions radiomics

model ($mean\ AUC = 0.71$), and the 21-subregions radiomics model ($mean\ AUC = 0.71$). This implied that a physiological segmentation of brain tumor volume into three subregions (WT, TC, and EC) played a pivotal role in the overall survival classification of brain gliomas. The 21-subregions radiomics model reported most stable predictions ($RSD = 1.39$), across six segmentation schemes, compared to the 6-subregions radiomics model ($RSD = 1.48$), the WT radiomics model ($RSD = 1.52$), and the 3-subregions radiomics model ($RSD = 1.99$). The stability of the 21-subregions radiomics model and 6-subregions radiomics model, over the 3-subregions radiomics model, is attributed to the sole dependence on the segmentation of WT volume which is more accurately generated by CNNs compared to TC and EC subregions.

Our failure analysis with the WT radiomics model, the 3-subregions radiomics model, the 6-subregions radiomics model, and the 21-subregions radiomics model revealed that 18 (distinct) subjects were misclassified by at least one radiomic model for a majority of segmentation networks. We found that the Hausdorff distance metric could be used to explain the afore-mentioned phenomena. More specifically, we focused on the HD-95 metric for WT segmentation which is common to the three radiomic models. Our analysis showed that the mean HD-95 metric (for WT segmentation), across six segmentation networks, for 13 correctly classified subjects (by majority of segmentation schemes) was $HD_{avg}^{WT} = 2.52 \pm 0.22$ and for 18 misclassified subjects was $HD_{avg}^{WT} = 5.92 \pm 1.17$. Moreover, 8 (out of 16) subjects which were misclassified by all radiomic models had large segmentation errors ($HD_{avg}^{WT} = 7.09 \pm 1.32$). This empirically demonstrated that strong predictive performance on overall survival classification of brain gliomas requires accurate segmentation of brain tumor volume with small segmentation errors.

We also found that most subjects that failed on at least one radiomics model were short-term survivors (8 subjects). Short-term survivors are typically associated with aggressive and heterogeneous tumor expressions [19] and, hence, one needs to augment the current feature set with appropriate measures of tumor heterogeneity for improved

classification. Our analysis also revealed that the WT radiomics model, the 6-subregions radiomics model, and the 21-subregions radiomics model simultaneously misclassified 11 (out of 12) subjects. This is attributed to the common requirement of an accurate segmentation of WT volume for feature extraction and classification.

8

Identification of robust features and evaluation of their impact on OS prediction

Identifying stable radiomic features is an important step for the translation of radiomics based approaches into clinical setting, because radiomic features are affected by several factors such as differences in image acquisition protocol, image reconstruction, and tumor segmentation. In this chapter, we present the experimental methodology followed to evaluate the robustness (in terms of stability) of radiomic features extracted from automatic segmentation algorithms and evaluate the impact of stable features on survival prediction, using the standard BraTS dataset of MRI scans as input data. Unlike in the previous study, we extracted a significantly larger number of features and performed feature selection on subset of radiomics features stable to variations in segmentations. For the segmentation, we used the same set of segmentation algorithms we used in Chapter

7, augmented with two additional segmentation networks: DMRes 3D CNN [81] and Pereira 2D U-Net [131].

All the implementation work was done in Python 3.6 using the following open-source packages: scikit-learn [127], N3 bias field correction [164], ANTs [7], PyRadiomics¹ [165], Pandas [113], ICC², OCCC³ and STAPLE fusion⁴ [142].

8.1 Experimental methodology

8.1.1 Data

As in the previous experiment, we made use of the publicly available BraTS 2020 dataset of 3D multiparametric MRI scans [115, 11, 10, 12]. A detailed description of the data is available in section 7.1.1. In this experiment, we used 2 of the data cohorts we used in the previous experiment: the Training cohort of 118 subjects, and Testing cohort A, comprising of 31 subjects.

8.1.2 Preprocessing

Preprocessing of the 3D MRI scans for each subject included skull-stripping, affine registration to the SRI24 template, resampling to an isotropic $1 \times 1 \times 1 \text{ mm}^3$ resolution, N3 bias correction, and mean-variance normalization [11, 151, 26].

8.1.3 Brain tumor segmentation

Manual segmentations of tumor subregions are already provided for the training cohort by BraTS challenge organizers. For the testing cohort, we automatically generated the segmentation of the brain tumor volume using the seven state-of-the-art CNNs discussed

¹<https://pyradiomics.readthedocs.io>

²<https://github.com/Mind-the-Pineapple/ICC>

³<https://rdr.io/cran/epiR/src/R/epi.occc.R>

⁴<https://github.com/FETS-AI/LabelFusion>

Table 8.1: Configuration and hyperparameters of the seven CNNs used for automatic segmentation of brain tumor volume (data provided by Syed. Talha Bukhari).

| Network | DeepMedicRes | Dong 2D U-Net | Wang 2.5D CNN | Isensee 3D U-Net | Pereira 2D U-Net | HDC-Net | E_1D_3 3D U-Net |
|--|---|---|--|--|---|---|---|
| Architecture | Dual-pathway (multi-scale) 3D CNN | 2D U-Net | Three 2.5D Anisotropic CNNs (W-Net, T-Net, and E-Net) in cascade | 3D U-Net with Deep supervision | 3D-2D U-Net in cascade | 2.5D U-Net | 3D U-Net |
| Activation | P-ReLU | ReLU | P-ReLU | Leaky-ReLU (0.01) | ReLU | ReLU | Leaky-ReLU (0.01) |
| Batch size | 10 | 10 | 5 (same for three CNNs in cascade) | 2 | 3D U-Net: 1 2D U-Net: 10 | 8 | 2 |
| Initialization | He-normal | He-normal | Truncated Normal | He-normal | He-normal | He-normal | He-normal |
| Input size / Output size | $52^2/9^3$ | $240^2/240^2$ | W-Net: $19 \times 144^2/11 \times 144^2$ T-Net: $19 \times 64^2/11 \times 64^2$ E-Net: $19 \times 64^2/11 \times 64^2$ | $128^3/128^3$ | 3D: $128^3/32^3$ 2D: $126^2/32^2$ | $128^3/128^3$ | $96^3/96^3$ |
| Learning Rate policy^a | Polynomial Decay (batch-wise) $\eta_0 = 10^{-4}$ $\eta_{end} = 10^{-7}$ $\gamma = 1.2$ | Polynomial Decay (batch-wise) $\eta_0 = 10^{-4}$ $\eta_{end} = 10^{-7}$ $\gamma = 1.2$ | Constant (10^{-3}) | Polynomial decay (epoch-wise) $\eta_0 = 0.01$ $\gamma = 0.9$ | Constant (5×10^{-5}) | Polynomial decay (epoch-wise) $\eta_0 = 10^{-3}$ $\gamma = 0.9$ | Polynomial decay (epoch-wise) $\eta_0 = 10^{-2}$ $\gamma = 0.9$ |
| Optimizer | Adam | Adam | Adam | SGD + Nesterov (0.99) | Adam | Adam (AMSGrad variant) | SGD + Nesterov (0.99) |
| Loss Function | Soft Dice | Soft Dice | Soft Dice | Soft Dice + Cross Entropy | Cross Entropy | Generalized Soft Dice | Soft Dice + Cross Entropy |
| Regularization | $L_1 (10^{-6})$, $L_2 (10^{-4})$ and Dropout (0.5) | – | $L_2 (10^{-7})$ | $L_2 (3 \times 10^{-5})$ | $L_2 (10^{-5})$, Spatial Dropout 2D and 3D: (0.05) | $L_2 (10^{-5})$ | $L_2 (10^{-6})$ |
| Total Training iterations (Gradient-Decent updates) | 100k (1000 epochs) | 50k (100 epochs) | 20k (per-network) | 250k (1000 epochs) | 3D: 100k (100 epochs) 2D: 30k (300 epochs) | 37.35k (900 epochs) | 125k (500 epochs) |
| # Parameters | 2.8 million | 34.5 million | W-Net: 0.21 million T-Net: 0.21 million E-Net: 0.20 million | 31.2 million | 3D: 0.68 million 2D: 1.64 million | 0.29 million | 34.9 million |
| Training Time^b | ~15 hours | ~110 hours | W-Net (single-view): ~84 hours T-Net (single-view): ~84 hours E-Net (single-view): ~20 hours | ~101 hours | 3D: ~12.5 hours 2D: ~23 hours | ~110 hours | ~48 hours |
| Test-time Augmentation | ✓ | ✓ | ✗ | ✓ | ✓ | ✓ | ✓ |
| Morphological Post-processing | Morphological closing, cluster thresholding | Morphological closing, cluster thresholding | ✗ | ✗ | Morphological closing, cluster thresholding | ✗ | ✓ |
| NOTES: ^a For definition of variables consult Table 1 in (Bukhari and Moly-ud-Din, 2021 [155]). ^b Please note that training time also depends on the GPU system used for training. HDC-Net was trained on a dual-GPU system whereas remaining CNNs were trained on a single-GPU system. | | | | | | | |

in detail in section 6.5 (DeepMedicRes, Dong 2D U-Net, Wang 2.5D CNN, Isensee 3D U-Net, Pereira 2D U-Net, HDC-Net, and E_1D_3 3D U-Net) after having trained them on the BraTS 2020 training data.

We also employed the STAPLE fusion method [142] to fuse the segmentation labels obtained from DeepMEDicRes, Dong 2D U-Net, Wang 2.5D CNN, Isensee 3D U-

Net, Pereira 2D U-Net HDC-Net, and E_1D_3 3D U-Net.

The segmentations were performed on a system with 64 GB RAM, and an NVIDIA RTX 2080Ti 11 GB GPU using the Tensorflow framework. Configuration and hyperparameters for the seven segmentation architectures are presented in Table 8.1.

8.1.4 Radiomic feature extraction

We extracted multi-regional and multi-modal radiomic features using the publicly available PyRadiomics software package [165]. More specifically, features were extracted for three overlapping tumor subregions (i.e., WT, TC, and EC) across four 3D MRI sequences (i.e., T1, T1-ce, T2, and FLAIR). For each tumor subregion and 3D MRI sequence, 98 features were extracted including 13 shape, 17 first-order, and 67 texture features. Moreover, first-order and texture features were extracted from the original 3D MRI image, wavelet filtered images (8 wavelet bands), and Laplacian of Gaussian filtered images ($\sigma = 1, 3$). Two additional shape features (volume and surface area) were extracted from the whole brain. In total, 11,129 radiomic features were extracted for each subject: 13 shape features \times 3 tumor subregions + 2 whole brain features + 84 first-order and texture features \times 4 channels \times 11 image filtering schemes (including original images) \times 3 tumor subregions. Table 8.2 lists the names of all radiomic features extracted for this study. For the training cohort, radiomic features were extracted from 3D mpMRI scans using manual segmentation of (overlapping) tumor subregions provided with the BraTS 2020 cohort. For the testing cohort, radiomic features were extracted using predicted segmentation maps from eight segmentation schemes (seven CNNs and the STAPLE-fusion method) discussed in in Section 6.5.

Radiomic features from the training cohort were inspected for outliers and NaNs. Outliers were identified with a scaled version of MAD as follows: $MAD_{scaled}(f_i) = cmedian(|f_i - f^-|)$ where $c = -\frac{1}{\{\sqrt{2}\theta^{-1}3/2\}}$ and θ^{-1} is the inverse complementary error function [97]. Every feature value $f_i^k > 3MAD_{scaled}(f_i)$ was labeled as an outlier and

replaced by the mean of the remaining feature values. The NaNs in each feature vector (f_i) were also replaced by the same mean value. Finally, z-score normalization was performed for each feature vector (f_i) by subtracting the mean and dividing by the standard deviation. Radiomic features from the testing cohort were also corrected for NaNs followed by z-score normalization. For each feature vector in the testing cohort, NaNs were replaced by the mean values used (to replace outliers and NaNs) in the training cohort and z-score normalization was performed using the mean and standard deviation computed on the training cohort.

Table 8.2: The list of radiomic features extracted for each subject in the training and testing cohorts.

| Feature Type | Feature Name | Total |
|--|---|-------|
| First order features | Energy, Entropy, Minimum, Maximum, 10th percentile, 90th percentile, Mean, Median, Interquartile Range, Range, MAD, rMAD, RMS, Skewness, Kurtosis, Variance, Uniformity. | 17 |
| Shape features | | |
| Multi-regional features | Mesh Volume, Surface Area, Surface Area to Volume Ratio, Sphericity, Maximum 3D Diameter, Maximum 2D Diameter (Slice), Maximum 2D Diameter (Column), Maximum 2D Diameter (Row), Major Axis Length, Minor Axis Length, Least Axis Length, Elongation, Flatness. | 13 |
| Total Brain features | Mesh Volume, Surface Area | 2 |
| Texture features | | |
| GLCM | Autocorrelation, Joint Average, Cluster Prominence, Cluster Shade, Cluster Tendency, Contrast, Correlation, Difference Average, Difference Entropy, Difference Variance, Joint Energy, Joint Entropy, IMC1, IMC2, MCC, IDMN, IDN, Inverse Variance, Maximum Probability, Sum Entropy, Sum of Squares. | 21 |
| GLRLM | Short Run Emphasis, Long Run Emphasis, Gray Level Non-Uniformity Normalized, Run Length Non-Uniformity Normalized, Run Percentage, Gray Level Variance, Run Variance, Run Entropy, Low Gray Level Run Emphasis, High Gray Level Run Emphasis, Short Run Low Gray Level Emphasis, Short Run High Gray Level Emphasis, Long Run Low Gray Level Emphasis, Long Run High Gray Level Emphasis | 14 |
| GLSZM | Small Area Emphasis, Large Area Emphasis, Gray Level Non-Uniformity, Size-Zone, Non-Uniformity Normalized, Zone Percentage, Gray Level Variance, Zone Variance, Zone Entropy, Low Gray Level Zone Emphasis, High Gray Level Zone Emphasis, Small Area Low Gray Level Emphasis, Small Area High Gray Level Emphasis, Large Area Low Gray Level Emphasis, Large Area High Gray Level Emphasis. | 14 |
| GLDM | Small Dependence Emphasis, Large Dependence Emphasis, Gray Level Non-Uniformity, Dependence Non-Uniformity Normalized, Gray Level Variance, Dependence Variance, Dependence Entropy, Low Gray Level Emphasis, High Gray Level Emphasis, Small Dependence Low Gray Level Emphasis, Small Dependence High Gray Level Emphasis, Large Dependence Low Gray Level Emphasis, Large Dependence High Gray Level Emphasis. | 13 |
| NGTDM | Coarseness, Contrast, Busyness, Complexity, Strength | 5 |
| Total Features | | 99 |
| NOTES: GLCM, gray-level co-occurrence matrix; GLRLM, gray-level run length matrix; GLSZM, gray-level size zone matrix; NGTDM, neighborhood gray-tone difference matrix; MAD, Mean Absolute Deviation; rMAD, Robust Mean Absolute Deviation; RMS, Root Mean Squared; IMC, Informational Measure of Correlation; IDMN, Inverse Difference Moment Normalized; MCC, Maximal Correlation Coefficient; ID, Inverse Difference. | | |

8.1.5 Stability analysis

We used the intra-class correlation coefficient (ICC) [150] and the overall concordance correlation coefficient (OCCC) [15] to quantitatively measure the robustness of radiomic features across the seven (independent) segmentation schemes we used. ICC is widely used to measure reliability of radiomic features across multiple raters [99, 67, 156]. In our study, we used ICC(2,1), which assumes that the seven state-of-the-art CNNs are sampled from a wider pool of deep segmentation networks reporting strong performance on brain tumor segmentation [150]. This is a reasonable assumption since the top performing methods are variants of an encoder-decoder architecture with distinct architectural and training hyperparameters. OCCC is another frequently used measure of agreement of radiomic features across multiple raters [99, 80] and is computed as a weighted average of all pairwise concordance correlation coefficients [94]. Unlike ICC, OCCC incorporates both the degree of agreement and disagreement by assigning higher weights to pairs of raters whose measurements have higher variances and larger mean differences [15]. In the spirit of domain adaptation, we selected a subset of radiomic features for training which are robust to variations in segmentations on the testing cohort. More specifically, the automatically generated brain tumor segmentations were considered as seven independent raters for reliability analysis. We synergistically used ICC and OCCC to select a subset of robust features from the original pool of 11129 radiomic features. A radiomic feature was considered to be robust if $ICC \geq 0.95$ and $OCCC \geq 0.95$. Reliability scores (ICC and OCCC) computed on the external testing cohort were only used to identify a pool of robust features for training a radiomic model, on the training cohort, which exhibits strong generalizability on the novel dataset.

8.1.6 Feature selection

To select an optimal subset of informative and discriminatory radiomic features, we employed a two-stage process. In stage 1, noninformative features with median absolute deviation (MAD) of zero were eliminated. In stage 2, a subset of discriminatory features was selected using one of the following feature selection methods:

Minimum Redundancy, Maximum Relevance (MRMR) [45]: MRMR is an information-theoretic approach of selecting a subset of minimally redundant features, F_i , quantified by a small average pairwise Pearson correlation $c(F_i, F_j)$ for all $1 \leq i, j \leq |F|$, which are strongly associated with response variables (Y), quantified by a large F-statistic $F(F_i, Y)$.

Recursive Feature Elimination with SVM (RFE-SVM) [137]: RFE starts by training an SVM classifier with the complete set of features and eliminates the one with the lowest feature importance score. This process is repeated on the reduced set of features until the required number of features are reached. RFE is superior to the forward feature selection approach as every feature is considered in the selection process.

8.1.7 OS prediction: model training and inference

Since this study aimed to determine whether stability analysis for the identification of robust features would lead to improved OS prediction performance compared to classification without stability analysis, we performed model training both with and without taking into consideration the results of stability analysis.

In the first case (with stability analysis), the original set of features was first reduced to a subset of robust features (via stability filtering as outlined in Section 8.1.5) followed by further reduction to a subset of informative and discriminatory features. The obtained subset of stable and discriminatory features was used to train fifty ($n = 50$) random forest classifiers with random initialization.

In the second case (without stability analysis), the original set of features was reduced to a subset of informative features (via MAD filtering) and discriminatory features (via one of the two feature selection methods outlined in section 8.1.6). The obtained subset of discriminatory features was used to train fifty ($n = 50$) random forest classifiers with random initialization.

Our hypothesis is that incorporating robust features in model training will enhance generalizability of the models on novel datasets. The original set of radiomic features (11,129 features) was augmented with a clinical feature, Age, before model training.

Hyperparameters of the random forest classifier were set as follows:

- RF classifier: $no_of_estimators = 200$, $max_features = auto$, $class_weight = balanced$, $criterion = gini$

Hyperparameters for the feature selection methods were set as follows:

- MRMR: $n_selected_features = N$
- RFE-SVM: $n_selected_features = N$, $kernel = linear$, $step = 1$

For model training, we explored using an optimal subset of features of varying cardinality including $N \in \{10, 15, 20, 25, 30, 40, 50, 70, 100\}$.

For the inference phase, a soft voting method was adopted to unify the outputs of 50 random forest classifiers (with uniform weighting scheme) and generate a single prediction (OS: short-term vs medium-term vs long-term).

To compute the predictive power of each radiomic feature, a single feature at a time was used to train a random forest classifier for classification. The uAUC of each feature was an average over 100 iterations of randomized and stratified splitting of the training cohort (70%-30% split).

8.1.8 Feature set reduction using prior information

In addition to the previously described feature selection approach, we explored an alternative feature selection approach based on exploiting prior information on feature robustness available in the literature.

Many studies in the literature report that shape features are the most predictive features for the OS classification task [132, 136, 2, 22, 55, 130, 124]. In particular, Suter et al. [156] investigated the robustness of different feature categories using 125 perturbations and reported that shape and location features are the most robust features for the OS classification task.

The first subset of features we selected was thus based on the list of robust features identified by Suter et al. [156]. We extracted these shape features from the overlapping whole tumor (WT) and tumor core (TC) subregions and from the non-overlapping peritumoral edema (PTE), non-enhancing core (NEC), and enhancing core (ENC) tumor subregions.

A second subset of features consisted in the spatial features discussed in section 7.1.5.

The third subset of features was based on the work of Pérez-Beteta et al. [132], who identified two contrast enhancement geometry (CEG) features that showed good predictive power for tumor geometry.

Finally, two additional shape features that are extracted from the whole brain (volume and surface area) were added to our set.

In total, using prior information for feature selection, we obtained 73 features for each subject: 13 shape features \times 5 tumor subregions, 4 spatial features, 2 CEG features, and 2 whole brain features.

8.1.9 Evaluation

Performance of the eight considered segmentation schemes was quantified using Dice Similarity Coefficient (DSC) [44] and Hausdorff distance metric (HD-95) [73]. The eight segmentation schemes were ranked based on the Final Ranking Score (FRS) and statistical significance (of ranking) was calculated using random permutation test [24]. Predictive performance of the radiomic models was quantified using area under the receiver operating curve (AUC). Stability of the radiomic models was quantified with relative standard deviation (RSD) calculated as a ratio of standard deviation to the mean of AUC. A lower value of RSD corresponds to higher stability of the radiomic model.

8.2 Results

8.2.1 Clinical characteristics

Clinical characteristics of the training and testing cohort are presented in table 7.1.

8.2.2 Segmentation algorithm performance

Table 8.3 summarizes the performance of the eight segmentation schemes for the testing dataset (125 subjects). The best overall segmentation performance for each tumor subregion corresponds to the highest Dice score (DSC) and lowest Hausdorff distance (HD-95). A high DSC implies that the predicted segmentation map has a high degree of overlap with the (ground truth) manual segmentation map. Low HD-95 implies that the predicted segmentation map has a low amount of voxel-wise segmentation error. We used Final Ranking Score (FRS) to unify the 7 segmentation performance metrics (i.e., DSC and HD-95 scores for three overlapping subregions each) for each subject in the testing dataset [24].

Table 8.3: Performance of the considered segmentation schemes on testing cohort (125 subjects).

| Segmentation Network | Dice Similarity Coefficient % | | | Hausdroff Distance (mm) | | | FRS |
|--------------------------------------|-------------------------------|-------------|-------------|-------------------------|--------------|--------------|-----|
| | WT | TC | ENC | WT | TC | ENC | |
| DMRes 3D CNN | 88.7 ± 12.3 | 78.1 ± 25.8 | 71.6 ± 31.6 | 8.97 ± 17.0 | 17.7 ± 57.3 | 32.3 ± 96.0 | 5** |
| Dong 2D U-Net | 89.6 ± 7.2 | 77.7 ± 23.7 | 71.0 ± 29.4 | 5.45 ± 7.7 | 11.4 ± 34.5 | 37.3 ± 105.0 | 7** |
| Wang 2.5D CNN | 88.1 ± 13.0 | 77.4 ± 25.3 | 75.2 ± 28.3 | 11.1 ± 20.8 | 13.67 ± 36.4 | 29.0 ± 91.2 | 6** |
| Isensee 3D U-Net | 90.5 ± 8.1 | 84.5 ± 16.4 | 76.9 ± 27.9 | 4.41 ± 5.99 | 8.65 ± 34.4 | 32.6 ± 100.9 | 1 |
| Pereira 2D U-Net | 87.7 ± 11.9 | 69.5 ± 30.2 | 67.0 ± 32.1 | 13.9 ± 23.3 | 22.85 ± 51.0 | 45.5 ± 108.8 | 8** |
| HDC-Net | 89.6 ± 10.3 | 93.1 ± 18.5 | 77.5 ± 27.2 | 7.5 ± 33.5 | 12.4 ± 47.5 | 32.3 ± 100.9 | 3** |
| E_1D_3 3D U- Net | 90.6 ± 6.4 | 82.7 ± 19.9 | 76.4 ± 27.6 | 5.8 ± 10.2 | 10.8 ± 35.9 | 22.96 ± 79.9 | 4** |
| STAPLE Fusion | 90.4 ± 7.4 | 82.9 ± 19.3 | 74.8 ± 28.8 | 5.3 ± 9.4 | 12.3 ± 47.4 | 30.6 ± 96.1 | 2 |

Note: ** indicates that the segmentation network is ranked significantly lower () in comparison to the top ranked method Isensee 3D U-Net (FRS = 1)

In terms of FRS, Isensee 3D U-Net and STAPLE fusion method were ranked first and second, respectively, with no significant difference between them ($p = 0.49$). However, the STAPLE fusion method and Isensee 3D U-Net were ranked significantly higher ($p < 0.001$) in comparison to the remaining six CNNs for brain tumor segmentation. Best overall segmentation performance for the WT and TC subregions were reported by E_1D_3 3D U-Net ($DSC = 90.6 \pm 6.4\%$ and $HD = 5.8 \pm 10.2mm$) and ($DSC = 82.7 \pm 19.9\%$ and $HD = 10.8 \pm 35.9mm$) respectively. No segmentation scheme reported best overall segmentation performance for the EC subregion. Predicted segmentation maps from HDC-Net maximally overlapped with the (ground-truth) manual segmentations ($DSC = 77.5 \pm 27.2\%$) but with large voxel-wise segmentation errors ($HD = 32.3 \pm 100.9mm$). On the contrary, E_1D_3 3D U-Net yielded predicted segmentation maps with (relatively) lowest voxel-wise segmentation errors ($HD = 22.9 \pm 79.9mm$).

8.2.3 Stability analysis results

From the original set of 11,129 features, extracted from the testing cohort, we first removed noninformative features (identified with $MAD=0$) and features not influenced by automatic segmentation, which includes Total Brain features (see feature list table) and Age. This reduced the original set to 11045 radiomic features.

From the pool of 11045 features, we then extracted the subset of radiomic features

that were stable to variations in the considered segmentations, by synergistically using the ICC and OCCC. More specifically, the subset consisted of radiomic features with an $ICC \geq 0.95$ and $OCCC \geq 0.95$ across the seven segmentation schemes. Figure 8.1 summarizes the number of stable features obtained for each feature category by stability filtering using ICC and OCCC independently.

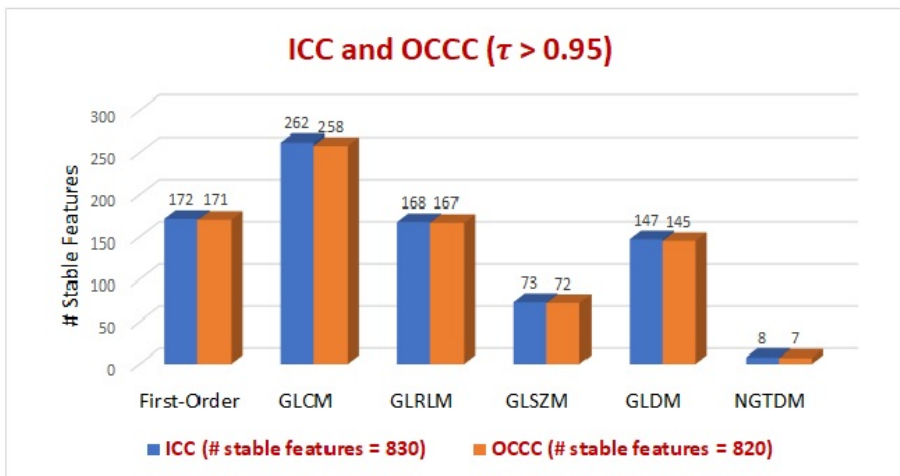


Figure 8.1: Distribution of stable features across different feature categories.

Stability filtering with ICC only ($\tau \geq 0.95$) yielded 830 stable features with the following statistics: Feature Category – 172 first-order features, 658 texture features, MRI Sequence – 153 features from FLAIR sequence, 299 features from T1ce sequence, 233 features from T1 sequence, and 145 features from T2 sequence, and Tumor Subregion – 801 features from WT region and 29 features from TC and zero feature from ENC subregions. Stability filtering with OCCC only ($\tau \geq 0.95$) yielded 820 stable features with the following statistics: Feature Category – 171 first-order features, 649 texture features, MRI Sequence – 151 features from FLAIR sequence, 294 features from T1ce sequence, 230 features from T1 sequence, and 145 features from T2 sequence, and Tumor Subregion – 791 features from WT region and 29 features from TC and zero feature from ENC subregions.

Stability filtering with both ICC and OCCC ($\tau \geq 0.95$) resulted in a subset of 820 stable radiomic features with the following statistics: Feature Category – 171 first-order features, 649 texture features, and zero shape feature, MRI Sequence – 151 features from FLAIR sequence, 249 features from T1ce sequence, 230 features from T1 sequence, and 145 features from T2 sequence, and Tumor Subregion – 791 features (96.5%) from WT region and 29 features (3.5%) from TC and zero feature from ENC subregions.

Our results from stability filtering revealed that the 820 highly stable radiomic features were: (1) predominantly texture features (79.1%), (2) mainly extracted from the WT region (96.5%), and (3) largely representing the FLAIR and T1ce sequences (58.4%).

The resulting set of highly stable radiomic features was augmented with Total Brain features (2) and Age to yield a subset of 823 features for prospective feature selection.

8.2.4 Overall survival classification

Feature selection

We employed one of the two feature selection methods – MRMR, and RFE-SVM – to obtain an optimal subset of discriminatory features for the underlying radiomic task. The size (or cardinality) of the optimal subset of features was controlled by prior setting the number of features (N), in MRMR and RFE-SVM. After feature selection, ten features were selected as shown in Table 8.4

Without stability filtering: the original set of radiomic and clinical features (11130) from the training cohort was first reduced to an informative subset of features via MAD filtering (11058 features) followed by a further reduction to a subset of discriminatory features:

(a) MRMR: We found that superior predictive power was obtained for an optimal subset of 10 features including Age, 2 first-order, and 7 texture features. Shape features were not selected. The statistics of the selected features were: mean OCCC 0.66 ± 0.3 ,

and mean uAUC 0.56 ± 0.04 . No stable features (ICC and OCCC ≥ 0.95) were selected with MRMR.

Table 8.4: A summary of features selected to build the model for the overall survival classification task.

| FCQ based MRMR- without stability filtering | | | | | | |
|---|--|-----------|--------|----------|-----------------|------|
| No. | Selected feature name | Type | Region | Modality | AUC \pm std | OCCC |
| f1 | GLRLM_Low Gray Level Run Emphasis | Texture | EN | T1ce | 0.57 ± 0.05 | 0.26 |
| f2 | GLRLM_Short Run High Gray Level Emphasis | Texture | WT | T2 | 0.56 ± 0.05 | 0.79 |
| f3 | GLCM_MCC | Texture | EN | FLAIR | 0.60 ± 0.06 | 0.68 |
| f4 | Skewness | Intensity | TC | T1 | 0.51 ± 0.05 | 0.36 |
| f5 | GLCM_Maximum Probability | Texture | WT | T2 | 0.62 ± 0.06 | 0.85 |
| f6 | Kurtosis | Intensity | EN | T2 | 0.60 ± 0.06 | 0.72 |
| f7 | GLCM_Imc2 | Texture | TC | T1ce | 0.62 ± 0.05 | 0.79 |
| f8 | GLRLM_Run Entropy | Texture | TC | FLAIR | 0.49 ± 0.06 | 0.86 |
| f9 | GLRLM_Short Run Low Gray Level Emphasis | Texture | EN | T1ce | 0.54 ± 0.05 | 0.27 |
| f10 | Age | Clinical | - | - | 0.54 ± 0.06 | 1 |
| FCQ based MRMR- with stability filtering | | | | | | |
| No. | Selected feature name | Type | Region | Modality | AUC \pm std | OCCC |
| f1 | GLCM_Contrast | Texture | WT | T1ce | 0.51 ± 0.06 | 0.98 |
| f2 | GLSZM_Zone Entropy | Texture | WT | T1 | 0.50 ± 0.06 | 0.97 |
| f3 | GLRLM_Run Variance | Texture | WT | T2 | 0.51 ± 0.05 | 0.96 |
| f4 | Uniformity | Intensity | TC | T1 | 0.57 ± 0.05 | 0.95 |
| f5 | GLRLM_Long Run High Gray Level Emphasis | Texture | WT | T1ce | 0.57 ± 0.06 | 0.96 |
| f6 | GLRLM_Long Run Emphasis | Texture | WT | T2 | 0.51 ± 0.05 | 0.97 |
| f7 | GLDM_Dependence Entropy | Texture | WT | T1 | 0.53 ± 0.05 | 0.99 |
| f8 | Robust Mean Absolute Deviation | Intensity | WT | T1ce | 0.52 ± 0.04 | 0.95 |
| f9 | GLCM_Inverse Variance | Texture | WT | T1ce | 0.45 ± 0.06 | 0.95 |
| f10 | Age | Clinical | - | - | 0.54 ± 0.06 | 1 |
| RFE-SVM without stability filtering | | | | | | |
| No. | Selected feature name | Type | Region | Modality | AUC \pm std | OCCC |
| f1 | GLCM_Difference Variance | Texture | EN | T1 | 0.54 ± 0.06 | 0.76 |
| f2 | GLRLM_Low Gray Level Run Emphasis | Texture | EN | T1ce | 0.57 ± 0.05 | 0.26 |
| f3 | GLRLM_Short Run Low Gray Level Emphasis | Texture | EN | T1ce | 0.54 ± 0.05 | 0.27 |
| f4 | GLRLM_Long Run Low Gray Level Emphasis | Texture | EN | T2 | 0.49 ± 0.05 | 0.56 |
| f5 | GLDM_Small Dependence High Gray Level Emphasis | Texture | TC | T1 | 0.54 ± 0.06 | 0.82 |
| f6 | GLRLM_Run Variance | Texture | TC | T1 | 0.52 ± 0.06 | 0.90 |
| f7 | Kurtosis | Intensity | TC | T2 | 0.52 ± 0.06 | 0.67 |
| f8 | Robust Mean Absolute Deviation | Intensity | TC | T2 | 0.58 ± 0.06 | 0.84 |
| f9 | 10Percentile | Intensity | WT | FLAIR | 0.54 ± 0.06 | 0.94 |
| f10 | GLRLM_Long Run Low Gray Level Emphasis | Texture | WT | T1 | 0.54 ± 0.07 | 0.53 |
| RFE-SVM with stability filtering | | | | | | |
| No. | Selected feature name | Type | Region | Modality | AUC \pm std | OCCC |
| f1 | GLCM_Joint Energy | Texture | WT | FLAIR | 0.47 ± 0.06 | 0.95 |
| f2 | GLDM_Large Dependence Emphasis | Texture | WT | FLAIR | 0.48 ± 0.06 | 0.95 |
| f3 | GLCM_Imc2 | Texture | WT | FLAIR | 0.45 ± 0.05 | 0.96 |
| f4 | GLCM_Joint Energy | Texture | WT | FLAIR | 0.49 ± 0.05 | 0.97 |
| f5 | GLSZM_Small Area High Gray Level Emphasis | Texture | WT | T1ce | 0.54 ± 0.05 | 0.97 |
| f6 | Uniformity | Intensity | TC | T1 | 0.57 ± 0.05 | 0.96 |
| f7 | GLRLM_Run Entropy | Texture | WT | T1 | 0.42 ± 0.06 | 0.99 |
| f8 | GLDM_Dependence Variance | Texture | WT | T2 | 0.55 ± 0.05 | 0.98 |
| f9 | GLDM_Large Dependence High Gray Level Emphasis | Texture | WT | T2 | 0.49 ± 0.06 | 0.96 |
| f10 | Age | Clinical | - | - | 0.54 ± 0.06 | 1 |

(b) RFE-SVM: We found that superior predictive power was obtained for an optimal subset of 10 features including 3 first-order, and 7 texture features. Shape features were not selected. The statistics of the selected features were: mean OCCC 0.66 ± 0.2 , and mean uAUC 0.54 ± 0.02 . No stable features ($\text{ICC and OCCC} \geq 0.95$) were selected with RFE-SVM.

With stability filtering: Post identification of an augmented subset of 823 stable radiomic features on the testing cohort, the corresponding feature labels were used to extract radiomic features from the training cohort. The resulting augmented subset of stable radiomic features (including Total Brain features and Age) was reduced to an informative subset of features via MAD filtering followed by further reduction to a subset of discriminatory features:

(a) MRMR: We found that superior predictive power was obtained for an optimal subset of 10 features including Age, 2 first-order, and 7 texture features. The statistics of the selected features were: mean OCCC 0.97 ± 0.02 , and mean AUC 0.52 ± 0.03 .

(b) RFE-SVM: We found that superior predictive power was obtained for an optimal subset of 10 features including Age, 1 first-order, and 8 texture features. The statistics of the selected features were: mean OCCC 0.97 ± 0.01 , and mean AUC 0.5 ± 0.05 .

Performance evaluation

Table 8.5 summarizes the predictive performance of the MRMR and RFE-SVM feature selection methods, with and without stability filtering, across the eight considered segmentation schemes (seven CNNs and STAPLE-fusion) using AUROC as quantitative measures. The robustness of radiomic models was quantified with Relative Standard Deviation (RSD) of AUROCs. To reiterate, learning a radiomic model requires a segmentation scheme (for volume of interest), feature reduction and selection pipeline, and a classifier.

Without stability filtering: The average predictive performance of the two feature

selection methods, across eight segmentation schemes, were as follows: MRMR – AUC 0.55 ± 0.06 and RFE-SVM – AUC 0.47 ± 0.04 . Isensee 3D U-Net showed strong predictive power for short-term survivors ($AUC = 0.7$) with features selected by MRMR and for long-term survivors ($AUC = 0.74$) with features selected by RFE-SVM feature selection method. The stability of the model was 10.4 with MRMR and 9.4 with RFE-SVM as measured with RSD, across the eight segmentation methods.

With stability filtering: The average predictive performance of the two feature selection methods, across eight segmentation schemes, were as follows: MRMR – AUC 0.91 ± 0.01 and RFE-SVM – AUC 0.91 ± 0.02 . Short-term and long-term predictive performance greatly improved for the seven segmentation schemes (CNNs), across two feature selection methods. Pereira 2D U-Net showed strong predictive power for short-term survivors ($AUC = 0.78$), and Wang 2.5D CNN for long-term survivors ($AUC = 0.79$), with features selected by the MRMR feature selection method. The stability of the model was 2.4 with MRMR and 2.2 with RFE-SVM as measured with RSD, across the eight segmentation methods.

Table 8.5: Model performance on the eight considered segmentation schemes on the testing cohort (31 subjects)

| Feature Selection Method | Stability filter status | Number of features | DeepMedicRes 3D CNN | Dong 2D U-Net | Wang 2.5D U-Net | Isensee 3D U-Net | Pereira 2D U-Net | HDC-Net | EID3 3D U-Net | STAPLE Fusion |
|--------------------------|-------------------------|--------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|----------------------------|
| FCQ based MRMR | True | 10 | 0.69 (0.74, 0.32, 0.75) | 0.66 (0.72, 0.31, 0.74) | 0.72 (0.76, 0.32, 0.79) | 0.68 (0.72, 0.32, 0.75) | 0.70 (0.78, 0.31, 0.75) | 0.69 (0.75, 0.35, 0.75) | 0.69 (0.72, 0.32, 0.77) | 0.71 (0.75, 0.35, 0.76) |
| | False | 10 | 0.48 (0.56, 0.49, 0.51) | 0.57 (0.61, 0.52, 0.55) | 0.46 (0.62, 0.48, 0.51) | 0.56 (0.7, 0.44, 0.58) | 0.59 (0.61, 0.45, 0.54) | 0.59 (0.68, 0.44, 0.53) | 0.65 (0.64, 0.54, 0.64) | 0.53 (0.66, 0.45, 0.58) |
| RFE-SVM | True | 10 | 0.59 (0.62, 0.38, 0.63) | 0.56 (0.63, 0.3, 0.59) | 0.57 (0.61, 0.33, 0.61) | 0.58 (0.63, 0.27, 0.63) | 0.58 (0.64, 0.24, 0.61) | 0.59 (0.65, 0.45, 0.59) | 0.57 (0.63, 0.35, 0.62) | 0.61 (0.65, 0.67, 0.61) |
| | False | 10 | 0.42 (0.56, 0.6, 0.7) | 0.43 (0.58, 0.4, 0.64) | 0.41 (0.55, 0.6, 0.64) | 0.49 (0.59, 0.49, 0.74) | 0.52 (0.64, 0.44, 0.6) | 0.51 (0.55, 0.62, 0.63) | 0.54 (0.51, 0.55, 0.49) | 0.47 (0.56, 0.52, 0.72) |

8.2.5 Effect of feature selection based on prior information

Stability analysis

From the original set of 73 features selected based on prior information, we first removed non-informative features (identified with $MAD=0$) and features not influenced by automatic segmentation, which includes Total Brain features and Age. This reduced the

original set to 71 radiomic features.

We then extracted a subset of radiomic features (from the pool of 71 features) which were stable to variations in segmentation. More specifically, highly stable radiomic features were identified with an $ICC \geq 0.90$ and $OCCC \geq 0.90$ across the seven segmentation schemes.

Stability filtering with ICC only ($\tau \geq 0.90$) yielded the following stable features: 3 PTE features (PTE_MeshVolume, PTE_surfaceArea, PTE_LeastAxisLength), 2 ENC (ENC_MeshVolume, ENC_SurfaceArea), 1 WT (WT_MeshVolume) and 3 spatial features (WT Centroid Coordinates (x, y, z)).

The obtained set of 9 highly stable radiomic features were finally augmented with the 2 Total Brain features and Age to yield a subset of 12 features for prospective feature selection.

Feature selection

We employed the two previously mentioned feature selection methods (MRMR and RFE-SVM) to obtain an optimal subset of discriminatory features for the overall survival classification task.

Without stability filtering: The original set of radiomic and clinical features (74), from the training cohort, was first reduced to an informative subset of features via MAD filtering (74 features) followed by further reduction to a subset of discriminatory features:

(a) MRMR: We found that superior predictive power was obtained for an optimal subset of 5 features including Age, 3 shape features (2 PTE and 1 WT) and 1 spatial feature. The statistics of the selected features were: mean $OCCC 0.797 \pm 0.13$, and mean $uAUC 0.55 \pm 0.006$. No stable features (ICC and $OCCC \geq 0.90$) were selected with MRMR.

(b) RFE-SVM: We found that superior predictive power was obtained for an optimal subset of 5 features including Age, 3 shape features (1 PTE, 1 TC, 1 NEC feature), and 1 spatial feature. The statistics of the selected features were: mean OCCC 0.79 ± 0.14 , and mean uAUC 0.55 ± 0.009 . No stable features (ICC and OCCC ≥ 0.90) were selected with RFE-SVM.

With stability filtering: Post identification of an augmented subset of 12 stable radiomic features on the testing cohort, the corresponding feature labels were used to extract radiomic features from the training cohort. The resulting augmented subset of stable radiomic features (including Total Brain features and Age) was reduced to an informative subset of features via MAD filtering followed by a further reduction to a subset of discriminatory features:

(a) MRMR: We found that superior predictive power was obtained for an optimal subset of 5 features including Age, 2 shape features (2 PTE), and 1 spatial and 1 total brain feature. The statistics of the selected features were: mean OCCC 0.96 ± 0.04 , and mean AUC 0.54 ± 0.02 .

(b) RFE-SVM: We found that superior predictive power was obtained for an optimal subset of 5 features including Age, 1 shape feature (PTE), and 2 spatial and 1 total brain feature. The statistics of the selected features were: mean OCCC 0.97 ± 0.03 , and mean AUC 0.54 ± 0.02 .

Performance evaluation

Table 8.6 summarizes the predictive performance of the two feature selection methods, with and without stability filtering, across the eight considered segmentation schemes using AUROC as quantitative measures. The robustness of radiomic models was quantified with Relative Standard Deviation (RSD) of AUROCs.

Without stability filtering: The average predictive performance of the two feature selection methods, across eight segmentation schemes, were as follows: MRMR – AUC 0.55 ± 0.01 and RFE-SVM – AUC 0.68 ± 0.02 . The stability of the model was 2.2 with MRMR and 2.9 with RFE-SVM as measured with RSD, across the eight segmentation methods.

With stability filtering: The average predictive performance of the two feature selection methods, across eight segmentation schemes, were as follows: MRMR – AUC 0.78 ± 0.01 and RFE-SVM – AUC 0.79 ± 0.009 . The stability of the model was 1.7 with MRMR and 1.2 with RFE-SVM as measured with RSD, across the eight segmentation methods.

Table 8.6: Model performance with the eight considered segmentation schemes on testing cohort (31 subjects) with priori selected features

| Feature Selection Method | Stability filter status | Number of features | DeepMedicRes 3D CNN | Dong 2D U-Net | Wang 2.5D U-Net | Isensee 3D U-Net | Pereira 2D U-Net | HDC-Net | E1D3 3D U-Net | STAPLE Fusion |
|--------------------------|-------------------------|--------------------|----------------------------|--------------------------|--------------------------|--------------------------|----------------------------|--------------------------|---------------------------|--------------------------|
| FCQ based MRMR | True | 5 | 0.77 (0.76,0.60,0.75) | 0.79 (0.79,0.51,0.79) | 0.80 (0.76,0.58,0.82) | 0.76 (0.72,0.46,0.80) | 0.80 (0.78,0.50,0.79) | 0.79 (0.75,0.55,0.80) | 0.77 (0.72,0.51,0.80) | 0.79 (0.75,0.57,0.79) |
| | False | 5 | 0.56 (0.51, 0.63, 0.60) | 0.57 (0.56,0.68,0.54) | 0.54 (0.51,0.57,0.59) | 0.56 (0.55,0.51,0.63) | 0.54 (0.5,0.76,0.54) | 0.53 (0.45,0.56,0.56) | 0.54 (0.50,0.64,0.57) | 0.56 (0.55,0.57,0.59) |
| RFE-SVM | True | 5 | 0.79 (0.81,0.50, 0.78) | 0.80 (0.86,0.48,0.68) | 0.80 (0.82,0.50,0.79) | 0.78 (0.81,0.45,0.80) | 0.81 (0.88,0.48,0.78) | 0.80 (0.82,0.50,0.80) | 0.78 (0.83, 0.45,0.81) | 0.79 (0.82,0.5,0.80) |
| | False | 5 | 0.71 (0.67, 0.77, 0.68) | 0.70 (0.70,0.54,0.65) | 0.72 (0.66,0.62,0.70) | 0.67 (0.69,0.38,0.69) | 0.67 (0.69, 0.49, 0.63) | 0.66 (0.65,0.52,0.65) | 0.66 (0.65,0.47,0.63) | 0.68 (0.66,0.57,0.63) |

8.3 Discussion

The translation of radiomics features into the clinical setting suffers from problems of reproducibility. Potential sources of variation in radiomics features such as image acquisition parameters, reconstruction algorithms, and software framework have already been evaluated by [156, 67, 166]. However, another source of variation is segmentation of ROI. Segmentation-induced variability due to CNN segmentation methods has not been explored in the literature. In this study, we quantified the stability of radiomic features across eight segmentation methods-seven CNNs and one STAPLE -fusion method. The seven CNNs include three 3D CNNs - Isensee 3D U-Net, E_1D_3 3D U-Net and DeepMedi-

cRes - two 2.5D CNNs - Wang 2.5D CNN and HDC-Net, two 2D CNNs, Dong 2D U-Net and Pereira 2D U-Net. Feature stability was calculated using segmentations generated for 125 subjects of BraTS 2020 testing cohort. In addition, we investigated whether the stability information could be used to improve the predictive performance of the random forest classifier for the prediction of overall survival. For comparison, the random forest classifier was trained with discriminatory features (via RFE-SVM, MRMR) alone and with both the stable and discriminatory features.

We used ICC and OCCC as measures of stability to evaluate the robustness of radiomic features in different segmentation networks. A few studies in the literature have used ICC [161, 67, 166] and OCCC [80] for radiomic feature reproducibility. We found no insight into the choice of stability measure. In our study, we used the intersection of ICC and OCCC. We found 820 stable features where (ICC and OCCC ≥ 0.95), the choice of threshold is random because we did not find a standard method for threshold selection in the literature.

Feature reproducibility differed between feature categories and for tumor subregions. We found that texture and first-order features were highly stable features. Shape features were the least stable features. The robustness of each feature reflects how much a small change in segmentation affects the feature value. One possible reason for this is that the intensity differences between tumor region and background are not very large and are less affected by tumor region segmentation. However, for shape features, a slight change in volume also has a strong effect on other features because they are derived from the volume feature. For different segmentation networks, we have different tumor sizes and thus different volumes. The results of our stability filtering show that the features most affected by segmentation are shape features. Overall, 80% of the stable features are from the WT region, which could be due to the fact that WT is a large contiguous region and is less affected by the segmentation algorithms.

We found that the predictive performance of the random forest model with the stable

and discriminatory features for predicting OS is much higher than that of the model trained with discriminatory features only. The stability-based AUCs outperformed the discriminatory AUCs in the eight segmentation schemes with two feature selection methods. Stable features and the MRMR method improved AUC ranges from 3.5% to 25.5%, while stable features and the RFE-SVM method improved AUC ranges from 3.8% to 16.3%. Models trained with both discriminatory and stable features increased the generalizability of the model. The E_1D_3 3D U-Net showed minimal improvement in AUC values of 3.5% and 3.8% (MRMR and RFE-SVM) as the WT volume segmentation achieved the highest Dice score ($DSC = 88.1 \pm 13.0\%$). However, we observed a large improvement in AUC values from 16% to 25.5% (MRMR and RFE-SVM) for Wang 2.5D CNN, with a large segmentation error of ($HD = 11.1 \pm 20.8mm$) for the WT region, so stable features bring more value.

In order to focus only on the shape features, we reduced the feature set (shape features, contrast enhancement geometry, spatial features, and clinical features). We performed another experiment by repeating the stability analysis, feature selection and then trained a classifier. The result shows higher AUC values by focusing only on the robust shape features.

9

Thesis conclusion and future work

Digital medical data collected through many different sources can effectively be used to investigate different medical conditions with the final aim of improving diagnosis and treatment of a condition and ultimately improve the life of affected patients. In this thesis, we worked with two medical datasets related to two different types of neurological disorders, motor control disorders (e.g., Parkinson's disease) and brain tumors, performing different types of analysis on the data to reach different goals.

In the first part of the thesis, we presented our work with a dataset containing the results of thousands of digital motor tests of the upper limbs taken by users of MotorBrain, a free and publicly available mobile application. Motor tests are used by neurologists to assess human motor performance and support the diagnosis of disorders affecting motor control. The first task we carried out on the data was to clean it using various criteria and solutions to remove incomplete and incorrect records. This is a necessary step when data is collected in the large using an unsupervised approach and

highlights the different types of issues that can occur in such a situation and their impact on the data. After identifying a set of measures that could best characterize the performance of users in the motor tests, we proceeded to analyze the data with a statistical approach with the goal of determining whether the data revealed typical patterns of human motor control performance such as the degradation of human motor performance that is typical of aging [95, 149]. The analysis focused on comparing motor performance across different age groups. Results show that the collected data reveal the expected patterns of human motor performance, thus providing evidence of the meaningfulness of the data and the appropriateness of the considered approach to motor performance data collection. At the same time, the results highlight potential problems that can emerge when data collection is performed in an unsupervised non-clinical setting. We then used machine learning techniques to automatically classify users based on their motor performance. Being limited to performance data of healthy individuals, we framed the classification problem as an age group identification problem. This could help neurologist to identify suspect cases at an early stage if a case does not behave in accordance with her age group normative behavior.

Future work on human motor performance assessment based on the MotorBrain dataset could move in different directions: (i) additional analysis can be performed on the data to explore other aspects of human motor performance such as gender differences, (ii) new measures can be identified that can better characterize the spatio-temporal behavior of users, e.g., based on a subdivision of trajectories into different parts, (iii) visualization tools can be developed that may support the visual exploration of the available data and complement the use of automatic analysis methods (e.g., by making it possible to quickly check the results of data cleaning activities), (iv) optimizations of the considered machine learning approaches and new approaches can be used to improve classification performance.

In the second part of the thesis, we focused on radiomics-based methods for overall

survival (OS) prediction of High Grade Glioma patients, using a standard dataset of 3D Multi-parametric Magnetic Resonance Imaging (MRI) scans. Our research goals were specifically related to two of the steps of the radiomic process: segmentation and feature selection. We first investigated the impact of different segmentation algorithms, five well-known Convolutional Neural Networks and the STAPLE-fusion method, on OS prediction based on four multiregional segmentation models, two physiology-based models (Whole Tumor (WT) and 3-subregions) and two atlas-guided anatomy-based models (6-subregions model and 21-subregions model). To do this, we applied the full radiomics process from preprocessing of the MRI scans to evaluation of the segmentation and prediction performance. In terms of segmentation performance, the Isensee 3D U-Net significantly outperformed the other CNNs based on dice similarity while the STAPLE fusion method was the best solution based on Hausdorff distance and second best for similarity. For OS prediction performance, the 3-subregions radiomic model proved to be the most predictive, but the 21-subregions and the 6-subregions model were the most stable across the six segmentation algorithms. Overall, we observed that good segmentation performance does not guarantee good radiomic performance and that short-term survivors are the most difficult to predict.

In a different experiment, we then evaluated the impact on OS prediction of selecting radiomic features based on stability analysis. To this end, we first measured the robustness of radiomic features across seven state-of-the-art (independent) segmentation methods based on Convolutional Neural Networks (CNNs), using the intra-class correlation coefficient (ICC) and the overall correlation coefficient (OCCC). We then employed two feature selection techniques, Minimum Redundancy, Maximum Relevance (MRMR) and Recursive Feature Elimination with SVM (RFE-SVM), to identify discriminatory features. Finally, we evaluated the effect of using robust radiomic features for OS classification by incorporating stability into feature selection methods, considering both stable features (via ICC and OCCC) and discriminatory features (via MRMR and RFE-SVM).

One limitation of our work, which is an often-encountered problem in clinical and translational imaging research, is that it is based on a small, even if standard, dataset. A large and balanced dataset would ideally help generalize the findings of our studies to diverse tumor manifestations and patient demographics. Different types of MRI datasets would also be useful. For example, Cepeda et al. [28] argue that perfusion and diffusion MRIs, along with structural MRIs, have the potential to improve outcome prediction for short-term survivors. While we employed different shape, volumetric, and spatial features for radiomics-based prediction of OS in brain gliomas, augmenting the current feature set with more stable and predictive features, capturing tumor heterogeneity and aggressiveness, may improve classification of short-term survivors in brain gliomas. Tumor heterogeneity is known to contribute to poor survival in high-grade gliomas [135]. To capture this heterogeneity, specific molecular markers and clinical information (gender, performance score, resection status) could be included for better performance [28]. Combining the radiomics-based prediction of OS with explainable artificial intelligence (XAI) would be interesting as well. The CNNs we used were trained using various combinations of Soft Dice and Cross Entropy loss functions. It would be interesting to see the impact of other loss functions, optimization schemes, and architectural engineering on segmentation accuracy and associated radiomics performance for OS classification in brain gliomas. Since our robustness study is based only on MR images, further studies with PET and CT may also add more value to the calculated results.

Bibliography

- [1] Rupal R Agravat and Mehul S Raval. Brain tumor segmentation and survival prediction. In *International MICCAI Brainlesion Workshop*, pages 338–348. Springer, 2019.
- [2] Rupal R Agravat and Mehul S Raval. 3d semantic segmentation of brain tumor for overall survival prediction. In *International MICCAI Brainlesion Workshop*, pages 215–227. Springer, 2020.
- [3] Agus Subhan Akbar, Chastine Fatichah, and Nanik Suciati. Modified mobilenet for patient survival prediction. In *International MICCAI Brainlesion Workshop*, pages 374–387. Springer, 2020.
- [4] Pablo Arias, Verónica Robles-García, Nelson Espinosa, Yoanna Corral, and Javier Cudeiro. Validity of the finger tapping test in parkinson’s disease, elderly and young healthy subjects: Is there a role for central fatigue? *Clinical Neurophysiology*, 123(10):2034–2041, 2012.
- [5] Teresa Arroyo-Gallego, María Jesus Ledesma-Carbayo, Alvaro Sánchez-Ferro, Ian Butterworth, Carlos S Mendoza, Michele Matarazzo, Paloma Montero, Roberto López-Blanco, Veronica Puertas-Martin, Rocio Trincado, et al. Detection of motor impairment in parkinson’s disease via mobile touchscreen typing. *IEEE Transactions on Biomedical Engineering*, 64(9):1994–2002, 2017.

- [6] Asra Aslam, Ekram Khan, and MM Sufyan Beg. Improved edge detection algorithm for brain tumor segmentation. *Procedia Computer Science*, 58:430–437, 2015.
- [7] Brian B Avants, Nick Tustison, Gang Song, et al. Advanced normalization tools (ants). *Insight j*, 2(365):1–35, 2009.
- [8] Leia B Bagesteiro and Robert L Sainburg. Handedness: dominant arm advantages in control of limb dynamics. *Journal of neurophysiology*, 88(5):2408–2421, 2002.
- [9] Leia B Bagesteiro and Robert L Sainburg. Nondominant arm advantages in load compensation during rapid elbow joint movements. *Journal of neurophysiology*, 90(3):1503–1513, 2003.
- [10] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin Kirby, John Freymann, Keyvan Farahani, and Christos Davatzikos. Segmentation labels and radiomic features for the pre-operative scans of the tcga-gbm collection. the cancer imaging archive. *Nat Sci Data*, 4:170117, 2017.
- [11] Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1):1–13, 2017.
- [12] Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

- [13] Sivakumar Balasubramanian, Alejandro Melendez-Calderon, Agnes Roby-Brami, and Etienne Burdet. On the analysis of movement smoothness. *Journal of neuroengineering and rehabilitation*, 12(1):1–11, 2015.
- [14] Subhashis Banerjee, Harkirat Singh Arora, and Sushmita Mitra. Ensemble of cnns for segmentation of glioma sub-regions with survival prediction. In *International MICCAI Brainlesion Workshop*, pages 37–49. Springer, 2019.
- [15] Huiman X Barnhart, Michael Haber, and Jingli Song. Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, 58(4):1020–1027, 2002.
- [16] Jérôme Barral and Bettina Debû. Hand and gender differences in the organization of aiming in 5-year-old children. *Neuropsychologia*, 40(2):152–161, 2002.
- [17] Stefan Bauer, Christof Seiler, Thibaut Bardyn, Philippe Buechler, and Mauricio Reyes. Atlas-based segmentation of brain tumor images using a markov random field-based tumor growth model and non-rigid registration. In *2010 annual international conference of the IEEE engineering in medicine and biology*, pages 4080–4083. IEEE, 2010.
- [18] Ivan Bautmans, Stijn Vantieghem, Ellen Gorus, Yuri-Reva Grazzini, Yves Fierens, Annelies Pool-Goudzwaard, and Tony Mets. Age-related differences in pre-movement antagonist muscle co-activation and reaction-time performance. *Experimental gerontology*, 46(8):637–642, 2011.
- [19] Niha Beig, Jay Patel, Prateek Prasanna, Virginia Hill, Amit Gupta, Ramon Correa, Kaustav Bera, Salendra Singh, Sasan Partovi, Vinay Varadan, et al. Radiogenomic analysis of hypoxia pathway is predictive of overall survival in glioblastoma. *Scientific reports*, 8(1):1–11, 2018.

- [20] Wenya Linda Bi and Rameen Beroukhim. Beating the odds: extreme long-term survival with glioblastoma, 2014.
- [21] R Blank, V Miller, and H Von Voss. Human motor development and hand laterality: a kinematic analysis of drawing movements. *Neuroscience Letters*, 295(3):89–92, 2000.
- [22] Vikas L Bommineni. Piecenet: A redundant unet ensemble. In *International MICCAI Brainlesion Workshop*, pages 331–341. Springer, 2020.
- [23] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
- [24] Syed Talha Bukhari and Hassan Mohy-ud Din. A systematic evaluation of learning rate policies in training cnns for brain tumor segmentation. *Physics in Medicine & Biology*, 66(10):105004, 2021.
- [25] Diedre Carmo, Leticia Rittner, and Roberto Lotufo. Multiattunet: Brain tumor segmentation and survival multitasking. In *International MICCAI Brainlesion Workshop*, pages 424–434. Springer, 2020.
- [26] Alexandre Carré, Guillaume Klausner, Myriam Edjlali, Marvin Lerousseau, Jade Briend-Diop, Roger Sun, Samy Ammari, Sylvain Reuzé, Emilie Alvarez Andres, Théo Estienne, et al. Standardization of brain mr images across machines and protocols: bridging the gap for mri-based radiomics. *Scientific reports*, 10(1):1–15, 2020.
- [27] Eric Carver, Chang Liu, Weiwei Zong, Zhenzhen Dai, James M Snyder, Joon Lee, and Ning Wen. Automatic brain tumor segmentation and overall survival prediction using machine learning algorithms. In *International MICCAI Brainlesion Workshop*, pages 406–418. Springer, 2018.

- [28] Santiago Cepeda, Angel Pérez-Nuñez, Sergio García-García, Daniel García-Pérez, Ignacio Arrese, Luis Jiménez-Roldán, Manuel García-Galindo, Pedro González, María Velasco-Casares, Tomas Zamora, et al. Predicting short-term survival after gross total or near total resection in glioblastomas by machine learning-based radiomic analysis of preoperative mri. *Cancers*, 13(20):5047, 2021.
- [29] S Cha. Update on brain tumor imaging: from anatomy to physiology. *American Journal of Neuroradiology*, 27(3):475–487, 2006.
- [30] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.
- [31] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021.
- [32] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.
- [33] Alessandro Cicerale, Elisabetta Ambron, Angelika Lingnau, and Raffaella I Rumiati. A kinematic analysis of age-related changes in grasping to use and grasping to move common objects. *Acta psychologica*, 151:134–142, 2014.
- [34] David A Cieslak, Nitesh V Chawla, and Aaron Striegel. Combating imbalance in network intrusion datasets. In *GrC*, pages 732–737, 2006.
- [35] Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, et al.

- The cancer imaging archive (tcia): maintaining and operating a public information repository. *Journal of digital imaging*, 26(6):1045–1057, 2013.
- [36] Erez James Cohen, Riccardo Bravi, Maria Angela Bagni, and Diego Minciocchi. Precision in drawing and tracing tasks: Different measures for different aspects of fine motor control. *Human movement science*, 61:177–188, 2018.
- [37] Jose L Contreras-Vidal, HansL Teulings, and GeorgeE Stelmach. Elderly subjects are impaired in spatial coordination in fine motor control. *Acta psychologica*, 100(1-2):25–35, 1998.
- [38] Thibaud P Coroller, Vishesh Agrawal, Elizabeth Huynh, Vivek Narayan, Stephanie W Lee, Raymond H Mak, and Hugo JW Aerts. Radiomic-based pathological response prediction from primary tumors and lymph nodes in nslc. *Journal of Thoracic Oncology*, 12(3):467–476, 2017.
- [39] AP Creagh, C Simillion, A Scotland, F Lipsmeier, C Bernasconi, S Belachew, J Van Beek, M Baker, C Gossens, M Lindemann, et al. Smartphone-based remote assessment of upper extremity function for multiple sclerosis using the draw a shape test. *Physiological measurement*, 41(5):054002, 2020.
- [40] Jessamyn Dahmen, Diane Cook, Robert Fellows, and Maureen Schmitter-Edgecombe. An analysis of a digital variant of the trail making test using machine learning techniques. *Technology and Health Care*, 25(2):251–264, 2017.
- [41] Christos Davatzikos, Saima Rathore, Spyridon Bakas, Sarthak Pati, Mark Bergman, Ratheesh Kalarot, Patmaa Sridharan, Aimilia Gastouniotti, Nariman Jahani, Eric Cohen, et al. Cancer imaging phenomics toolkit: quantitative imaging analytics for precision diagnostics and predictive modeling of clinical outcome. *Journal of medical imaging*, 5(1):011018, 2018.

- [42] Luciane Aparecida Pascucci Sande de Souza, Valdeci Carlos Dionísio, and Gil Lúcio Almeida. Multi-joint movements with reversal in parkinson’s disease: Kinematics and electromyography. *Journal of Electromyography and Kinesiology*, 21(2):376–383, 2011.
- [43] Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.
- [44] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.
- [45] Chris Ding and Hanchuan Peng. Minimum redundancy feature selection from microarray gene expression data. *Journal of bioinformatics and computational biology*, 3(02):185–205, 2005.
- [46] Hao Dong, Guang Yang, Fangde Liu, Yuanhan Mo, and Yike Guo. Automatic brain tumor detection and segmentation using u-net based fully convolutional networks. In *annual conference on medical image understanding and analysis*, pages 506–517. Springer, 2017.
- [47] ERlet al Dorsey, R Constantinescu, JP Thompson, KM Biglan, RG Holloway, K Kiebertz, FJ Marshall, BM Ravina, G Schifitto, A Siderowf, et al. Projected number of people with parkinson disease in the most populous nations, 2005 through 2030. *Neurology*, 68(5):384–386, 2007.
- [48] Charalampos Doukas, Thomas Pliakas, and Ilias Maglogiannis. Mobile healthcare information management utilizing cloud computing and android os. In *2010 An-*

- nual International Conference of the IEEE Engineering in Medicine and Biology*, pages 1037–1040. IEEE, 2010.
- [49] Natalia Dounskaia, Laetitia Fradet, Gyusung Lee, Berta C Leis, and Charles H Adler. Submovements during pointing movements in parkinson’s disease. *Experimental brain research*, 193(4):529–544, 2009.
- [50] Natalia Dounskaia, Arend WA Van Gemmert, Berta C Leis, and George E Stelmach. Biased wrist and finger coordination in parkinsonian patients during performance of graphical tasks. *Neuropsychologia*, 47(12):2504–2514, 2009.
- [51] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdenek Smékal, and Marcos Faundez-Zanuy. Analysis of in-air movement in handwriting: A novel marker for parkinson’s disease. *Computer methods and programs in biomedicine*, 117(3):405–411, 2014.
- [52] Peter Drotár, Jiří Mekyska, Irena Rektorová, Lucia Masarová, Zdeněk Smékal, and Marcos Faundez-Zanuy. Decision support framework for parkinson’s disease based on novel handwriting markers. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 23(3):508–516, 2014.
- [53] Amir Fallahi and Shahram Jafari. An expert system for detection of breast cancer using data preprocessing and bayesian network. *International Journal of Advanced Science and Technology*, 34:65–70, 2011.
- [54] Andriy Fedorov, Reinhard Beichel, Jayashree Kalpathy-Cramer, Julien Finet, Jean-Christophe Fillion-Robin, Sonia Pujol, Christian Bauer, Dominique Jennings, Fiona Fennessy, Milan Sonka, et al. 3d slicer as an image computing platform for the quantitative imaging network. *Magnetic resonance imaging*, 30(9):1323–1341, 2012.

- [55] Xue Feng, Nicholas J Tustison, Sohil H Patel, and Craig H Meyer. Brain tumor segmentation using an ensemble of 3d u-nets and overall survival prediction using radiomic features. *Frontiers in computational neuroscience*, 14:25, 2020.
- [56] Lucas Fidon, Sébastien Ourselin, and Tom Vercauteren. Generalized wasserstein dice score, distributionally robust deep learning, and ranger for brain tumor segmentation: Brats 2020 challenge. In *International MICCAI Brainlesion Workshop*, pages 200–214. Springer, 2020.
- [57] Paul M Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954.
- [58] Valeria Fonti and Eduard Belitser. Feature selection using lasso. *VU Amsterdam research paper in business analytics*, 30:1–25, 2017.
- [59] Jie Fu, Kamal Singhrao, Xinran Zhong, Yu Gao, Sharon X Qi, Yingli Yang, Dan Ruan, and John H Lewis. An automatic deep learning–based workflow for glioblastoma survival prediction using preoperative multimodal mr images: A feasibility study. *Advances in radiation oncology*, 6(5):100746, 2021.
- [60] Josep Garre-Olmo, Marcos Faúndez-Zanuy, Karmele López-de Ipiña, Laia Calvó-Perxas, and Oriol Turró-Garriga. Kinematic and pressure features of handwriting and drawing: preliminary results between patients with mild cognitive impairment, alzheimer disease and healthy controls. *Current Alzheimer research*, 14(9):960–968, 2017.
- [61] Mina Ghaffari, Arcot Sowmya, and Ruth Oliver. Brain tumour segmentation using cascaded 3d densely-connected u-net. *arXiv preprint arXiv:2009.07563*, 2020.
- [62] Robert J Gillies, Paul E Kinahan, and Hedvig Hricak. Radiomics: images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016.

- [63] Ali Gooya, Kilian M Pohl, Michel Bilello, Luigi Cirillo, George Biros, Elias R Melhem, and Christos Davatzikos. Glistr: glioma image segmentation and registration. *IEEE transactions on medical imaging*, 31(10):1941–1954, 2012.
- [64] Neil Grech, Theresia Dalli, Sean Mizzi, Lara Meilak, Neville Calleja, and Antoine Zrinzo. Rising incidence of glioblastoma multiforme in a well-defined population. *Cureus*, 12(5), 2020.
- [65] Philipp Gulde and Joachim Hermsdörfer. Smoothness metrics in complex movement tasks. *Frontiers in neurology*, 9:615, 2018.
- [66] Xiaoqing Guo, Chen Yang, Pak Lun Lam, Peter YM Woo, and Yixuan Yuan. Domain knowledge based brain tumor segmentation and overall survival prediction. In *International MICCAI Brainlesion Workshop*, pages 285–295. Springer, 2019.
- [67] Christoph Haarbuerger, Gustav Müller-Franzes, Leon Weninger, Christiane Kuhl, Daniel Truhn, and Dorit Merhof. Radiomics feature reproducibility under inter-rater variability in segmentations of ct images. *Scientific reports*, 10(1):1–10, 2020.
- [68] Theophraste Henry, Alexandre Carré, Marvin Lrousseau, Théo Estienne, Charlotte Robert, Nikos Paragios, and Eric Deutsch. Brain tumor segmentation with self-ensembled, deeply-supervised 3d u-net neural networks: a brats 2020 challenge solution. In *International MICCAI Brainlesion Workshop*, pages 327–339. Springer, 2020.
- [69] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

- [70] Rui Hua, Quan Huo, Yaozong Gao, Yu Sun, and Feng Shi. Multimodal brain tumor segmentation using cascaded v-nets. In *International MICCAI Brainlesion Workshop*, pages 49–60. Springer, 2018.
- [71] Kerry A Hubel, E William Yund, Timothy J Herron, and David L Woods. Computerized measures of finger tapping: reliability, malingering and traumatic brain injury. *Journal of Clinical and Experimental Neuropsychology*, 35(7):745–758, 2013.
- [72] Kit Huckvale, Samanta Adomaviciute, José Tomás Prieto, Melvin Khee-Shing Leow, and Josip Car. Smartphone apps for calculating insulin dose: a systematic assessment. *BMC medicine*, 13(1):1–10, 2015.
- [73] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.
- [74] Dimitrios Iakovakis, Stelios Hadjidimitriou, Vasileios Charisis, Sevasti Bostantzopoulou, Zoe Katsarou, and Leontios J Hadjileontiadis. Touchscreen typing-pattern analysis for detecting fine motor skills decline in early-stage parkinson’s disease. *Scientific reports*, 8(1):1–13, 2018.
- [75] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021.
- [76] Mobarakol Islam, V Jose, and Hongliang Ren. Glioma prognosis: Segmentation of the tumor and survival prediction using shape, geometric and clinical information. In *International MICCAI Brainlesion Workshop*, pages 142–153. Springer, 2018.

- [77] Mobarakol Islam and Hongliang Ren. Multi-modal pixnet for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 298–308. Springer, 2017.
- [78] Aravind Kailas, Chia-Chin Chong, and Fujio Watanabe. From mobile phones to personal wellness dashboards. *IEEE pulse*, 1(1):57–63, 2010.
- [79] Lorraine V Kalia and Anthony E Lang. Parkinson’s disease. *The Lancet*, 386(9996):896–912, 2015.
- [80] Jayashree Kalpathy-Cramer, Artem Mamomov, Binsheng Zhao, Lin Lu, Dmitry Cherezov, Sandy Napel, Sebastian Echegaray, Daniel Rubin, Michael McNitt-Gray, Pechin Lo, et al. Radiomics of lung nodules: a multi-institutional study of robustness and agreement of quantitative imaging features. *Tomography*, 2(4):430–437, 2016.
- [81] Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V Nori, Antonio Criminisi, Daniel Rueckert, and Ben Glocker. Deepmedic for brain tumor segmentation. In *International workshop on Brainlesion: Glioma, multiple sclerosis, stroke and traumatic brain injuries*, pages 138–149. Springer, 2016.
- [82] Konstantinos Kamnitsas, Christian Ledig, Virginia FJ Newcombe, Joanna P Simpson, Andrew D Kane, David K Menon, Daniel Rueckert, and Ben Glocker. Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. *Medical image analysis*, 36:61–78, 2017.
- [83] Anuraag R Kansal, S Torquato, GR Harsh Iv, EA Chiocca, and TS Deisboeck. Simulated brain tumor growth dynamics using a three-dimensional cellular automaton. *Journal of theoretical biology*, 203(4):367–382, 2000.
- [84] Po-Yu Kao, Thuyen Ngo, Angela Zhang, Jefferson W Chen, and BS Manjunath. Brain tumor segmentation and tractographic feature extraction from structural

- mr images for overall survival prediction. In *International MICCAI Brainlesion Workshop*, pages 128–141. Springer, 2018.
- [85] Seyed-Mahdi Khaligh-Razavi, Sina Habibi, Maryam Sadeghi, Haniye Marefat, Mahdijeh Khanbagi, Seyed Massood Nabavi, Elham Sadeghi, and Chris Kalafatis. Integrated cognitive assessment: speed and accuracy of visual processing as a reliable proxy to cognitive performance. *Scientific Reports*, 9(1):1–11, 2019.
- [86] Laura Klaming and Björn NS Vlaskamp. Non-dominant hand use increases completion time on part b of the trail making test but not on part a. *Behavior Research Methods*, 50(3):1074–1087, 2018.
- [87] Saskia Klenk, Doreen Reifegerste, and Rebecca Renatus. Gender differences in gratifications from fitness app use and implications for health interventions. *Mobile Media & Communication*, 5(2):178–193, 2017.
- [88] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Montreal, Canada, 1995.
- [89] Vasilis Kontis, James E Bennett, Colin D Mathers, Guangquan Li, Kyle Foreman, and Majid Ezzati. Future life expectancy in 35 industrialised countries: projections with a bayesian model ensemble. *The Lancet*, 389(10076):1323–1335, 2017.
- [90] C Kotsavasiloglou, Nikolaos Kostikis, Dimitrios Hristu-Varsakelis, and Marianthi Arnaoutoglou. Machine learning-based classification of simple drawing movements in parkinson’s disease. *Biomedical Signal Processing and Control*, 31:174–180, 2017.
- [91] Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):1–8, 2017.

- [92] Christoph Laske, Hamid R Sohrabi, Shaun M Frost, Karmele López-de Ipiña, Peter Garrard, Massimo Buscema, Justin Dauwels, Surjo R Soekadar, Stephan Mueller, Christoph Linnemann, et al. Innovative diagnostic tools for early detection of alzheimer’s disease. *Alzheimer’s & Dementia*, 11(5):561–578, 2015.
- [93] Andrius Lauraitis, Rytis Maskeliūnas, Robertas Damaševičius, and Tomas Krilavičius. A mobile application for smart computer-aided self-administered testing of cognition, speech, and motor impairment. *Sensors*, 20(11):3236, 2020.
- [94] I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.
- [95] Jonas SR Leversen, Monika Haga, and Hermundur Sigmundsson. From children to adults: motor performance across the life-span. *PloS one*, 7(6):e38830, 2012.
- [96] Oron Levin, Hakuei Fujiyama, Matthieu P Boisgontier, Stephan P Swinnen, and Jeffery J Summers. Aging and motor inhibition: a converging perspective provided by brain stimulation and imaging approaches. *Neuroscience & Biobehavioral Reviews*, 43:100–117, 2014.
- [97] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology*, 49(4):764–766, 2013.
- [98] Jundong Li, Kewei Cheng, Suhang Wang, Fred Morstatter, Robert P Trevino, Jiliang Tang, and Huan Liu. Feature selection: A data perspective. *ACM computing surveys (CSUR)*, 50(6):1–45, 2017.
- [99] Zhuoran Li, Huichuan Duan, Kun Zhao, Yanhui Ding, and Yuanjie Zheng. Stability of mri radiomic features of the hippocampus: An integrated analysis of test-

- retest variability. In *2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB)*, pages 140–144. IEEE, 2019.
- [100] Andy Liaw, Matthew Wiener, et al. Classification and regression by randomforest. *R news*, 2(3):18–22, 2002.
- [101] Geng-Cheng Lin, Wen-June Wang, Chung-Chia Kang, and Chuin-Mu Wang. Multispectral mr images segmentation based on fuzzy knowledge and modified seeded region growing. *Magnetic resonance imaging*, 30(2):230–246, 2012.
- [102] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016.
- [103] Yang Liu, Nitesh V Chawla, Mary P Harper, Elizabeth Shriberg, and Andreas Stolcke. A study in machine learning from imbalanced data for sentence boundary detection in speech. *Computer Speech & Language*, 20(4):468–494, 2006.
- [104] Zhihua Liu, Long Chen, Lei Tong, Feixiang Zhou, Zheheng Jiang, Qianni Zhang, Caifeng Shan, Yinhai Wang, Xiangrong Zhang, Ling Li, et al. Deep learning based brain tumor segmentation: a survey. *arXiv preprint arXiv:2007.09479*, 2020.
- [105] David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.
- [106] P Low. Merck manual. *Consumer Version-Overview of the Autonomic Nervous System* (<http://www.merckmanuals.com/home/brain-spinal-cord-andnerve-disorders/autonomic-nervous-system-disorders/overview-of-the-autonomicnervous-system>), 2013.

- [107] Bradley Christopher Lowekamp, David T Chen, Luis Ibáñez, and Daniel Blezek. The design of simpleitk. *Frontiers in neuroinformatics*, 7:45, 2013.
- [108] Zhengrong Luo, Zhongdao Jia, Zhimin Yuan, and Jialin Peng. Hdc-net: Hierarchical decoupled convolution network for brain tumor segmentation. *IEEE Journal of Biomedical and Health Informatics*, 25(3):737–745, 2020.
- [109] Luke Macyszyn, Hamed Akbari, Jared M Pisapia, Xiao Da, Mark Attiah, Vadim Pigrish, Yingtao Bi, Sharmistha Pal, Ramana V Davuluri, Laura Roccograndi, et al. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. *Neuro-oncology*, 18(3):417–425, 2015.
- [110] Padma R Mahant and Mark A Stacy. Movement disorders and normal aging. *Neurologic clinics*, 19(3):553–563, 2001.
- [111] Jaime Marti Asenjo and Alfonso Martinez-Larraz Solís. Mri brain tumor segmentation using a 2d-3d u-net ensemble. In *International MICCAI Brainlesion Workshop*, pages 354–366. Springer, 2020.
- [112] Richard McKinley, Micheal Rebsamen, Katrin Daetwyler, Raphael Meier, Piotr Radojewski, and Roland Wiest. Uncertainty-driven refinement of tumor-core segmentation using 3d-to-2d networks with label uncertainty. In *International MICCAI Brainlesion Workshop*, pages 401–411. Springer, 2020.
- [113] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.
- [114] Raphael Meier, Urspeter Knecht, Tina Loosli, Stefan Bauer, Johannes Slotboom, Roland Wiest, and Mauricio Reyes. Clinical evaluation of a fully-automatic segmentation method for longitudinal brain tumor volumetry. *Scientific reports*, 6(1):1–11, 2016.

- [115] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2014.
- [116] Georgia Mitsi, Enrique Urrea Mendoza, Benjamin D Wissel, Elena Barbopoulou, Alok K Dwivedi, Ioannis Tsoulos, Athanassios Stavrakoudis, Alberto J Espay, and Spyros Papapetropoulos. Biometric digital health technology for measuring motor function in parkinson’s disease: results from a feasibility and patient satisfaction study. *Frontiers in neurology*, 8:273, 2017.
- [117] Momina Moetesum, Imran Siddiqi, Nicole Vincent, and Florence Cloppet. Assessing visual attributes of handwriting for prediction of neurological disorders — a case study on parkinson’s disease. *Pattern Recognition Letters*, 121:19–27, 2019.
- [118] Hajar Moradmand, Seyed Mahmoud Reza Aghamiri, and Reza Ghaderi. Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma. *Journal of applied clinical medical physics*, 21(1):179–190, 2020.
- [119] Hieu T Nguyen, Tung T Le, Thang V Nguyen, and Nhan T Nguyen. Enhancing mri brain tumor segmentation with an additional classification network. In *International MICCAI Brainlesion Workshop*, pages 503–513. Springer, 2020.
- [120] Robert L Nussbaum and Christopher E Ellis. Alzheimer’s disease and parkinson’s disease. *New england journal of medicine*, 348(14):1356–1364, 2003.
- [121] World Health Organization. *Neurological disorders: public health challenges*. World Health Organization, 2006.
- [122] Quinn T Ostrom, Haley Gittleman, Gabrielle Truitt, Alexander Boscia, Carol Kruchko, and Jill S Barnholtz-Sloan. Cbtrus statistical report: primary brain and

- other central nervous system tumors diagnosed in the united states in 2011–2015. *Neuro-oncology*, 20(suppl_4):iv1–iv86, 2018.
- [123] Nikolaos Papanikolaou, Celso Matos, and Dow Mu Koh. How to develop a meaningful radiomic signature for clinical use in oncologic patients. *Cancer Imaging*, 20(1):1–10, 2020.
- [124] Bhavesh Parmar and Mehul Parikh. Brain tumor segmentation and survival prediction using patch based modified 3d u-net. In *International MICCAI Brainlesion Workshop*, pages 398–409. Springer, 2020.
- [125] Chintan Parmar, Patrick Grossmann, Johan Bussink, Philippe Lambin, and Hugo JWL Aerts. Machine learning methods for quantitative radiomic biomarkers. *Scientific reports*, 5(1):1–11, 2015.
- [126] Linda Partridge, Joris Deelen, and P Eline Slagboom. Facing up to the global challenges of ageing. *Nature*, 561(7721):45–56, 2018.
- [127] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.
- [128] Jan C Peeken, Matthew B Spraker, Carolin Knebel, Hendrik Dapper, Daniela Pfeiffer, Michal Devecka, Ahmed Thamer, Mohamed A Shouman, Armin Ott, Rüdiger von Eisenhart-Rothe, et al. Tumor grading of soft tissue sarcomas using mri-based radiomics. *EBioMedicine*, 48:332–340, 2019.
- [129] Linmin Pei, Spyridon Bakas, Arastoo Vossough, Syed MS Reza, Christos Davatzikos, and Khan M Iftekharuddin. Longitudinal brain tumor segmentation prediction in mri using feature and label fusion. *Biomedical signal processing and control*, 55:101648, 2020.

- [130] Linmin Pei, AK Murat, and Rivka Colen. Multimodal brain tumor segmentation and survival prediction using a 3d self-ensemble resnet. In *International MICCAI Brainlesion Workshop*, pages 367–375. Springer, 2020.
- [131] Sérgio Pereira, Adriano Pinto, Victor Alves, and Carlos A Silva. Brain tumor segmentation using convolutional neural networks in mri images. *IEEE transactions on medical imaging*, 35(5):1240–1251, 2016.
- [132] Julián Pérez-Beteta, Alicia Martínez-González, David Molina, Mariano Amo-Salas, Belén Luque, Elena Arregui, Manuel Calvo, José M Borrás, Carlos López, Marta Claramonte, et al. Glioblastoma: does the pre-treatment geometry matter? a postcontrast t1 mri-based study. *European radiology*, 27(3):1096–1104, 2017.
- [133] Michael Peters, Stian Reimers, and John T Manning. Hand preference for writing and associations with selected demographic and behavioral variables in 255,100 subjects: the bbc internet study. *Brain and cognition*, 62(2):177–189, 2006.
- [134] Nicole Porz, Simon Habegger, Raphael Meier, Rajeev Verma, Astrid Jilch, Jens Fichtner, Urspeter Knecht, Christian Radina, Philippe Schucht, Jürgen Beck, et al. Fully automated enhanced tumor compartmentalization: man vs. machine reloaded. *PLoS One*, 11(11):e0165302, 2016.
- [135] Prateek Prasanna, Jay Patel, Sasan Partovi, Anant Madabhushi, and Pallavi Tiwari. Radiomic features from the peritumoral brain parenchyma on treatment-naive multi-parametric mr imaging predict long versus short-term survival in glioblastoma multiforme: preliminary findings. *European radiology*, 27(10):4188–4197, 2017.
- [136] Elodie Puybureau, Guillaume Tochon, Joseph Chazalon, and Jonathan Fabrizio. Segmentation of gliomas and prediction of patient overall survival: a simple and

- fast procedure. In *International MICCAI Brainlesion Workshop*, pages 199–209. Springer, 2018.
- [137] Alain Rakotomamonjy. Variable selection using svm-based criteria. *Journal of machine learning research*, 3(Mar):1357–1370, 2003.
- [138] Rachael K Raw, Georgios K Kountouriotis, Mark Mon-Williams, and Richard M Wilkie. Movement control in older adults: does old age mean middle of the road? *Journal of experimental psychology: human perception and performance*, 38(3):735, 2012.
- [139] Ralph M Reitan. Validity of the trail making test as an indicator of organic brain damage. *Perceptual and motor skills*, 8(3):271–276, 1958.
- [140] Stefania Rizzo, Francesca Botta, Sara Raimondi, Daniela Origgi, Valentina Buscarino, Anna Colarieti, Federica Tomao, Giovanni Aletti, Vanna Zanagnolo, Maria Del Grande, et al. Radiomics of high-grade serous ovarian cancer: association between quantitative ct features, residual tumour and disease progression within 12 months. *European radiology*, 28(11):4849–4859, 2018.
- [141] Shannon D Robertson. Development of bimanual skill: The search for stable patterns of coordination. *Journal of motor behavior*, 33(2):114–126, 2001.
- [142] Torsten Rohlfing, Daniel B Russakoff, and Calvin R Maurer. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE transactions on medical imaging*, 23(8):983–994, 2004.
- [143] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

- [144] David A Rosenbaum. *Human motor control*. Academic press, 2009.
- [145] Abdulwahab Sahyoun, Karim Chehab, Osama Al-Madani, Fadi Aloul, and Assim Sagahyroon. Parknosis: Diagnosing parkinson’s disease using mobile phones. In *2016 IEEE 18th International Conference on e-Health Networking, Applications and Services (Healthcom)*, pages 1–6. IEEE, 2016.
- [146] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018.
- [147] Parita Sanghani, Beng Ti Ang, Nicolas Kon Kam King, and Hongliang Ren. Overall survival prediction in glioblastoma multiforme patients from volumetric, shape and texture features using machine learning. *Surgical oncology*, 27(4):709–714, 2018.
- [148] Lisa Scarpace, L Mikkelsen, T Cha, Sujaya Rao, Sangeeta Tekchandani, S Gutman, and D Pierce. Radiology data from the cancer genome atlas glioblastoma multiforme [tcga-gbm] collection. *The Cancer Imaging Archive*, 11(4):1, 2016.
- [149] Rachael D Seidler, Jessica A Bernard, Taritonye B Burutolu, Brett W Fling, Mark T Gordon, Joseph T Gwin, Youngbin Kwak, and David B Lipps. Motor control and aging: links to age-related brain structural, functional, and biochemical effects. *Neuroscience & Biobehavioral Reviews*, 34(5):721–733, 2010.
- [150] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.
- [151] John G Sled, Alex P Zijdenbos, and Alan C Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE transactions on medical imaging*, 17(1):87–97, 1998.

- [152] Sebastian Starke, Carlchristian Eckert, Alex Zwanenburg, Stefanie Speidel, Steffen Löck, and Stefan Leger. An integrative analysis of image segmentation and survival of brain tumour patients. In *International MICCAI Brainlesion Workshop*, pages 368–378. Springer, 2019.
- [153] Christoph Straube, Kerstin A Kessel, Stefanie Antoni, Jens Gempt, Bernhard Meyer, Juergen Schlegel, Friederike Schmidt-Graf, and Stephanie E Combs. A balanced score to predict survival of elderly patients newly diagnosed with glioblastoma. *Radiation Oncology*, 15(1):1–11, 2020.
- [154] Wen-Tsai Sung and Yen-Chun Chiang. Improved particle swarm optimization algorithm for android medical care iot using modified parameters. *Journal of medical systems*, 36(6):3755–3763, 2012.
- [155] Yannick Suter, Alain Jungo, Michael Rebsamen, Urspeter Knecht, Evelyn Herrmann, Roland Wiest, and Mauricio Reyes. Deep learning versus classical regression for brain tumor patient survival prediction. In *International MICCAI Brainlesion Workshop*, pages 429–440. Springer, 2018.
- [156] Yannick Suter, Urspeter Knecht, Mariana Alão, Waldo Valenzuela, Ekkehard Hewer, Philippe Schucht, Roland Wiest, and Mauricio Reyes. Radiomics for glioblastoma survival analysis in pre-operative mri: exploring feature robustness, class boundaries, and machine learning techniques. *Cancer Imaging*, 20(1):1–13, 2020.
- [157] Syed Talha Bukhari and Hassan Mohy-ud Din. E1d3 u-net for brain tumor segmentation: Submission to the rsna-asnr-miccai brats 2021 challenge. *arXiv e-prints*, pages arXiv–2110, 2021.
- [158] Jiliang Tang, Salem Alelyani, and Huan Liu. Feature selection for classification: A review. *Data classification: Algorithms and applications*, page 37, 2014.

- [159] Paweł Teisseyre. Ccnet: Joint multi-label classification and feature selection using classifier chains and elastic net regularization. *Neurocomputing*, 235:98–111, 2017.
- [160] Kiran T Thakur, Emiliano Albanese, Panteleimon Giannakopoulos, Nathalie Jette, Mattias Linde, Martin J Prince, Timothy J Steiner, Tarun Dua, et al. Neurological disorders. *Mental, Neurological, and Substance Use Disorders*, page 87, 2016.
- [161] Florent Tixier, Hyemin Um, Robert J Young, and Harini Veeraraghavan. Reliability of tumor segmentation in glioblastoma: impact on the robustness of mri-radiomic features. *Medical physics*, 46(8):3582–3591, 2019.
- [162] Gerard J Tortora and Bryan H Derrickson. *Principles of anatomy and physiology*. John Wiley & Sons, 2018.
- [163] WJ Triggs, R Calvanio, M Levine, RK Heaton, and KM Heilman. Predicting hand preference with performance on motor tasks. *Cortex*, 36(5):679–689, 2000.
- [164] Nicholas J Tustison, Brian B Avants, Philip A Cook, Yuanjie Zheng, Alexander Egan, Paul A Yushkevich, and James C Gee. N4itk: improved n3 bias correction. *IEEE transactions on medical imaging*, 29(6):1310–1320, 2010.
- [165] Joost JM Van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, Regina GH Beets-Tan, Jean-Christophe Fillion-Robin, Steve Pieper, and Hugo JW Aerts. Computational radiomics system to decode the radiographic phenotype. *Cancer research*, 77(21):e104–e107, 2017.
- [166] R Verma, VB Hill, V Statsevych, K Bera, R Correa, P Leo, M Ahluwalia, A Madabhushi, and P Tiwari. Stable and discriminatory radiomic features from the tumor and its habitat associated with progression-free survival in glioblastoma: A multi-institutional study. *American Journal of Neuroradiology*, 43(8):1115–1123, 2022.

- [167] Andrea Vianello, Luca Chittaro, Stefano Burigat, and Riccardo Budai. Motor-brain: a mobile app for the assessment of users' motor performance in neurology. *Computer methods and programs in biomedicine*, 143:35–47, 2017.
- [168] Max A Viergever, JB Antoine Maintz, Stefan Klein, Keelin Murphy, Marius Staring, and Josien PW Pluim. A survey of medical image registration—under review, 2016.
- [169] Stéphanie Vuillermot, Anina Pescatore, Lisa Holper, Daniel C Kiper, and Kynan Eng. An extended drawing test for the assessment of arm and hand function with a performance invariant for healthy subjects. *Journal of Neuroscience Methods*, 177(2):452–460, 2009.
- [170] Shuo Wang, Chengliang Dai, Yuanhan Mo, Elsa Angelini, Yike Guo, and Wenjia Bai. Automatic brain tumour segmentation and biophysics-guided survival prediction. In *International MICCAI Brainlesion Workshop*, pages 61–72. Springer, 2019.
- [171] Andreas Wibmer, Hedvig Hricak, Tatsuo Gondo, Kazuhiro Matsumoto, Harini Veeraraghavan, Duc Fehr, Junting Zheng, Debra Goldman, Chaya Moskowitz, Samson W Fine, et al. Haralick texture analysis of prostate mri: utility for differentiating non-cancerous prostate from prostate cancer and differentiating prostate cancers with different gleason scores. *European radiology*, 25(10):2840–2850, 2015.
- [172] SP Wise and R Shadmehr. Motor control. encyclopedia of the human brain, 2000.
- [173] Jonathan R Wood, Sylvan B Green, and William R Shapiro. The prognostic importance of tumor size in malignant gliomas: a computed tomographic scan study by the brain tumor cooperative group. *Journal of clinical oncology*, 6(2):338–343, 1988.

- [174] Shuang Wu, Jin Meng, Qi Yu, Ping Li, and Shen Fu. Radiomics-based machine learning methods for isocitrate dehydrogenase genotype prediction of diffuse gliomas. *Journal of cancer research and clinical oncology*, 145(3):543–550, 2019.
- [175] Hai Xu, Hongtao Xie, Yizhi Liu, Chuandong Cheng, Chaoshi Niu, and Yongdong Zhang. Deep cascaded attention network for multi-task brain tumor segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 420–428. Springer, 2019.
- [176] Hikari Yamashita. Right-and left-hand performance on the rey–osterrieth complex figure: A preliminary study in non-clinical sample of right handed people. *Archives of Clinical Neuropsychology*, 25(4):314–317, 2010.
- [177] Shuojue Yang, Dong Guo, Lu Wang, and Guotai Wang. Cascaded coarse-to-fine neural network for brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 458–469. Springer, 2020.
- [178] Yang Yang, Yu Han, Xintao Hu, Wen Wang, Guangbin Cui, Lei Guo, and Xin Zhang. An improvement of survival stratification in glioblastoma patients via combining subregional radiomics signatures. *Frontiers in neuroscience*, 15:559, 2021.
- [179] Yi Yang, Heng Tao Shen, Zhigang Ma, Zi Huang, and Xiaofang Zhou. L2, 1-norm regularized discriminative feature selection for unsupervised. In *Twenty-second international joint conference on artificial intelligence*, 2011.
- [180] Hongdou Yao, Xiaobing Zhou, and Xuejie Zhang. Automatic segmentation of brain tumor using 3d se-inception networks with residual connections. In *International MICCAI Brainlesion Workshop*, pages 346–357. Springer, 2018.
- [181] Yading Yuan. Automatic brain tumor segmentation with scale attention network. In *International MICCAI Brainlesion Workshop*, pages 285–294. Springer, 2020.

- [182] Habib Zaidi and Issam El Naqa. Pet-guided delineation of radiation therapy treatment volumes: a survey of image segmentation techniques. *European journal of nuclear medicine and molecular imaging*, 37(11):2165–2187, 2010.
- [183] Poonam Zham, Sridhar P Arjunan, Sanjay Raghav, and Dinesh K Kumar. Efficacy of guided spiral drawing in the classification of parkinson’s disease. *IEEE journal of biomedical and health informatics*, 22(5):1648–1652, 2017.
- [184] Dingwen Zhang, Guohai Huang, Qiang Zhang, Jungong Han, Junwei Han, and Yizhou Yu. Cross-modality deep feature learning for brain tumor segmentation. *Pattern Recognition*, 110:107562, 2021.
- [185] Guojing Zhao, Bowen Jiang, Jianpeng Zhang, and Yong Xia. Segmentation then prediction: A multi-task solution to brain tumor segmentation and survival prediction. In *International MICCAI Brainlesion Workshop*, pages 492–502. Springer, 2020.
- [186] Yuan-Xing Zhao, Yan-Ming Zhang, and Cheng-Lin Liu. Bag of tricks for 3d mri brain tumor segmentation. In *International MICCAI Brainlesion Workshop*, pages 210–220. Springer, 2019.
- [187] Alex Zwanenburg, Stefan Leger, Martin Vallières, and Steffen Löck. Image biomarker standardisation initiative. *arXiv preprint arXiv:1612.07003*, 2016.