université PARIS-SACLAY

Machine learning for genomics and imaging data integration applied to neuro-oncology

Machine learning pour l'intégration des données génomiques et d'imagerie appliquée à la neuro-oncologie

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 575 : electrical, optical, bio : physics and engineering (EOBE)
Spécialité de doctorat : Sciences de l'information et de la communication
Graduate School : Sciences de l'ingénierie et des systèmes. Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'unité de recherche **BAOBAB (Université Paris-Saclay, CEA, CNRS)**, sous la direction de **Vincent FROUIN**, Docteur, HDR et le co-encadrement de **Cathy PHILIPPE**, Ingénieure de Recherche

Thèse soutenue à Paris-Saclay, le 23 mars 2023, par

Hamza CHEGRAOUI

Composition du jury

Membres du jury avec voix délibérative

Arthur TENENHAUS Professeur, Université Paris-Saclay, Centrale-Supélec	Président
Céline LEFEBVRE Chercheuse, HDR, Servier, Computational Medicine	Rapporteur & Examinatrice
Laurent GUYON Chercheur, HDR, CEA Grenoble - Université de Grenoble Alpes	Rapporteur & Examinateur
Irène BUVAT Directrice de recherche, Université Paris Saclay, CNRS, Institut Curie	Examinatrice
Anaïs BAUDOT Directrice de recherche, Université Aix Marseille, CNRS, Marseille Medical Genetics	Examinatrice

THESE DE DOCTORAT
NNT : 2023UPAST040

université PARIS-SACLAY

ÉCOLE DOCTORALE
Physique et ingénierie : Electrons, Photons, Sciences du vivant (EOBE)

Titre : Machine learning pour l'intégration des données génomiques et d'imagerie appliquée à la neuro-oncologie

Mots clés :traitement d'images, apprentissage statistique, tumeurs cérébrales, intégration de données

Résumé : Le gliome infiltrant du tronc cérébral (Diffuse Intrinsic Pontine Glioma (DIPG)) est une tumeur cérébrale rare située dans le pons, principalement observé chez les enfants âgés de 5 à 7 ans. Elle est considérée comme l'une des tumeurs pédiatriques les plus agressives, avec un taux de survie inférieur à 10% au-delà des deux ans après le diagnostic et une médiane de survie globale inférieure à un an. Le DIPG est classé comme un gliome diffus de la ligne médiane (DMG), principalement caractérisé par une mutation K27M des gènes codant pour la protéine histone H3 et/ou une perte de la triméthylation H3K27 par surexpression de la protéine EZHIP. L'emplacement de la tumeur et ses altérations génomiques correspondantes fait du DIPG un type de tumeur complètement différent des autres tumeurs de haut grade.

Ce travail propose l'intégration des données d'imagerie avec les données génétiques afin de trouver des biomarqueurs. Dans un premier temps, nous nous intéressons à l'extraction des régions d'intérêt des images nécessaires pour une étude radiomique. Ensuite, nous proposons une procédure d'intégration des données multi-sources, qui prend en compte les graphes complexes d'interaction entre les gènes. Finalement, nous appliquons notre procédure sur les données disponibles afin de comparer ses performances avec d'autres modèles de la littérature et étudier l'apport de l'imagerie et du graphe aux données génétiques.

L'analyse radiomique nécessite des régions d'intérêt prédéfinies sur les images disponibles. Pour notre cohorte DIPG, la segmentation manuelle de la tumeur n'était pas disponible car elle ne fait pas partie de la routine clinique. De plus, aucune base de données spécifique n'a été créée pour entraîner des algorithmes d'apprentissage automatique classiques afin de délimiter automatiquement les régions tumorales. Cette étude s'est concentrée sur l'obtention de segmentations binaires pour le DIPG en utilisant uniquement les modalités FLAIR et T2w, à partir de modèles entraînés sur des gliomes de haut grade. Nous proposons de combiner différents modèles simples de détection et de segmentation pour obtenir des résultats de segmentation satisfaisants.

En parallèle, un modèle d'intégration multi-blocs prenant en compte des graphes complexes connus d'interactions entre les gènes a été développé et l'influence du graphe choisi sur la sélection des variables par le modèle a été étudiée. Nous proposons netSGCCA, un modèle combinant la Sparse Generalized Canonical Correlation Analysis (SGCCA) et la pénalité GraphNet. Nous avons appliqué notre modèle à l'ensemble de données multi-omiques TCGA-LGG (The Cancer Genome Atlas – Low Grade Glioma). Contrairement à Elastic-Net seul, la pénalité GraphNet est capable de sélectionner un ensemble raisonnable de gènes tout en offrant une interprétation biologique au niveau des voies biologiques et donc informative sur de potentielles cibles thérapeutiques. L'exemple sur l'ensemble de données TCGA-LGG montre la stabilité et la fiabilité de netSGCCA pour la sélection des variables d'intérêt.

Enfin, nous avons utilisé netSGCCA pour intégrer la radiomique et les données génétiques et l'avons appliqué à la tâche de prédiction de la survie. En raison de l'indisponibilité des données de survie sur notre cohorte DIPG, nous avons utilisé l'ensemble de données TCGA-LGG pour mener l'étude. Nous avons comparé netSGCCA avec d'autres approches de survie multi-blocs et des modèles construits en mono-blocs. La netSGCCA s'est révélée être un modèle robuste capable de sélectionner des gènes connus dans le gliome de bas-grade et d'autres interagissants dans des voies biologiques pertinentes. Le bloc radiomique n'a pas fourni d'information supplémentaire au modèle. Cependant, les caractéristiques radiomiques extraites de la modalité T2, en mono-bloc sont des prédicteurs performants, ce qui représenterait un espoir pour les patients avec DIPG qui n'ont pas de biopsie.

Title : Machine learning for genomics and imaging data integration applied to neuro-oncology
Keywords : Image processing, statistical learning, brain tumours, data integration

Abstract : Diffuse Intrinsic Pontine Glioma (DIPG) is a rare brain tumour located in the pons, mainly seen in children aged 5 to 7 years. It is considered one of the most aggressive paediatric tumours, with a survival rate of less than 10% beyond two years after diagnosis and a median overall survival of less than one year. DIPG is classified as a diffuse midline glioma (DMG), mainly characterized by a K27M mutation of the genes encoding the histone H3 protein and/or a loss of H3K27 trimethylation by overexpression of the EZHIP protein. The location of the tumour and its corresponding genomic alterations makes DIPG a completely different type of tumour than other high-grade gliomas.

This work proposes the integration of imaging data with genetic data in order to find biomarkers. First, we are interested in the extraction of the regions of interest of the images necessary for a radiomic study. Then, we propose a procedure for the integration of multi-source data, which takes into account the complex graphs of interaction between genes. Finally, we apply our procedure to the available data in order to compare its performance with other existing models and to study the contribution of imagery and the graph to genetic data.

Radiomic analysis requires predefined regions of interest on available images. For our DIPG cohort, manual tumour segmentation was not feasible. Moreover, no database was created to train classical machine learning algorithms to automatically delineate tumor regions. This study focused on obtaining binary segmentations for DIPG using only FLAIR and T2w modalities, using models trained on glioblastomas. Our proposition combines different simple detection and segmentation models to obtain satisfactory segmentation results.

In parallel, our work aims to build a multi-block integration model taking into account the intra-block correlation structure described in established complex graphs of gene-gene interactions (e.g. PathwayCommons). Moreover, our objective is to understand how the interaction graph influences the selection of variables. We propose netSGCCA, a model combining the Sparse Generalized Canonical Correlation Analysis (SGCCA) with the GraphNet penalty. We applied our model to the TCGA-LGG dataset. Unlike Elastic-Net alone, the GraphNet penalty was able to select a reasonable set of genes and gives an informative biological interpretation using biological pathway enrichment analysis. The example on the TCGA-LGG dataset shows the stability and reliability of netSGCCA for selecting variables of interest.

Finally, we used netSGCCA to integrate radiomics and genetic data and applied it to the survival prediction task. Due to the unavailability of survival data on our DIPG cohort, we used the TCGA-LGG dataset to conduct the study. We compared the results obtained with netSGCCA with other multi-block survival approaches and models built in mono-blocks. The netSGCCA has proven to be a robust model capable of selecting variables already linked to the pathology studied and interacting in relevant biological pathways. The addition of imageing did not improve the predictive ability of netSGCCA. However, the baseline results show that the radiomic features extracted from the T2 modality can be strong predictors.

Acknowledgements

I want to start my acknowledgements by thanking Doctor Jacques Grill and his team at the Gustave Roussy Institute. I deeply respect your commitment to paediatric glioma patients and your interest in various research angles. This thesis would not have been possible without your collaboration and the Imagine for Margo association, which funded this work.

I am also thankful to my thesis supervisors Vincent Frouin and Cathy Philippe. You gave me a unique opportunity to work on an interesting subject. The feedback I received always helped me improve and grow. Furthermore, I would like to thank all former and current members of the Brainomics team, who welcomed me among them and with whom conversation was always interesting. I would also thank Neurospin and all its members for allowing me to work in such a diverse research laboratory.

I would also like to thank the people that evaluated my PhD work: Céline Lefebvre, Laurent Guyon, Arthur Tenenhaus, Irène Buvat and Anaïs Baudot. Thank you for taking some of your precious time to review my work. Both the questions raised and the discussion were very insightful. It is a deep honour to have you as my jury members.

I would also acknowledge the profound impact of those around me during my whole life, especially my thesis preparation. My parents' support is critical for who I am today. I conclude by thanking my partner, Lina. Your curiosity towards my work and patience with me was a source of motivation.

Contents

List of Figures vii

List of Tables x

List of Algorithms xiii

1 Introduction 1

1.1 Open questions raised by multi-view datasets in rare tumours 4

1.2 Manuscript outline 6

I Materials and Methods 7

2 Notations and Definitions 9

2.1 Notations 10

2.2 Data description 10

2.2.1 Imaging Data 10

2.2.2 Genetic data 14

2.3 Method Definitions 14

2.3.1 Objective functions 15

2.3.2 Regularisation 15

2.3.3 Classical models 16

2.3.4 Metrics 19

3 Related Works 21

3.1 Radiomics 23

3.1.1 Classification of tumour types or subtypes using radiomics 23

3.1.2 Survival Prediction using radiomics 23

3.1.3 Reproducibility of radiomics 24

3.2 Brain tumour segmentation 26

3.2.1 Using custom features 26

3.2.2 Using Deep-learning . . . 27
3.2.3 Tackling issues with brain tumour segmentation . . . 28
3.3 Multi-block approaches . . . 29
3.3.1 Unsupervised models . . . 29
3.3.2 Supervised models . . . 31
3.4 Variable selection . . . 32
3.4.1 Sparsity-based penalties . . . 32
3.4.2 Graph based penalties . . . 33

4 Datasets Description and Preprocessing 37
4.1 Datasets Description . . . 39
4.1.1 Glioblastoma . . . 39
4.1.2 Lower Grade Gliomas . . . 40
4.1.3 DIPG . . . 41
4.1.4 Data availability . . . 41
4.2 Preprocessing . . . 42
4.2.1 Image preprocessing . . . 42
4.2.2 Radiomic extraction . . . 45
4.2.3 Mutation data preprocessing . . . 46

II Contributions 47

5 Brain Tumour Segmentation 49
5.1 Introduction . . . 51
5.2 Detection-segmentation combination strategy . . . 52
5.2.1 Combination strategies . . . 52
5.2.2 Final masking . . . 52
5.2.3 Input data . . . 53
5.2.4 Ensembling the inferences . . . 54
5.3 Reusing off-the-shelf networks . . . 54
5.3.1 You Only Look Once (YOLO) . . . 54
5.3.2 UNet and BB-UNet models . . . 56
5.3.3 Deepmedic . . . 56
5.4 Experimental designs . . . 57
5.5 Benchmark results . . . 58
5.5.1 Object-detection results . . . 58
5.5.2 Segmentation results on DIPG . . . 63
5.6 Discussion . . . 65
5.7 Conclusion . . . 67

6 Multiblock Integration Method 69
6.1 Regularized Generalized Canonical Correlation Analysis . . . 71

6.2 Sparse Generalized Canonical Correlation Analysis . . . 72
6.3 netSGCCA . . . 72
6.3.1 GraphNet . . . 73
6.3.2 Optimisation . . . 74
6.3.3 The proximal operator . . . 75
6.4 Application . . . 76
6.4.1 Simulated Data . . . 76
6.4.2 TCGA-LGG analysis design and results . . . 80
6.4.3 Discussion . . . 88
6.4.4 Conclusion . . . 89

7 Radio-genomics integration and survival prediction 91
7.1 Methodology . . . 93
7.1.1 Other Multi-block survival models from the state-of-the-art . . . 93
7.1.2 Experimental design . . . 94
7.2 Baseline results using all per-block available data . . . 97
7.2.1 Radiomics . . . 97
7.2.2 Somatic mutations . . . 100
7.2.3 RNA . . . 103
7.3 Comparison mono-block and multi-block approaches . . . 104
7.3.1 Mono-block . . . 104
7.3.2 Multi-block . . . 108
7.4 Discusion . . . 111
7.5 Conclusion . . . 112

8 Conclusion 113

A First Appendix 115

B Second Appendix 119

C Résumé en français 121
C.1 Introduction . . . 121
C.2 Contribution . . . 123
C.2.1 Chapitre 5 : Segmentation de la tumeur . . . 123
C.2.2 Chapitre 6 : Intégration multi-blocks . . . 124
C.2.3 Chapitre 7 : Intégration Radio-génomique . . . 124
C.3 Conclusion . . . 125

Bibliography 127

List of Figures

2.1 Acquired T2 signal. (a) As the spins realign with the magnetic field, their electrical signal decreases. The decrease is characterised by the T2 relaxation time. (b) The T2 relaxation time is tissue dependant. Acquiring the signal at TE gives the contrast between the different tissues and results in a T2-weighted image. 11

2.2 Acquired T1 signal. (a) As the spins realign with the magnetic field, their magnetism increases. The increase is characterised by the T1 relaxation time. (b) The T1 relaxation time is tissue dependant. Acquiring the signal at TR gives the contrast between the different tissues and results in a T1-weighted image. 11

2.3 A description of how the different texture features are computed. In an example 4×4 image ROI, three gray levels are represented by numerical values from 1 to 3. The red and blue annotations are used to highlight the different voxels used in order to compute some matrix values. To illustrate the different matrices: For the GLCM we only considered voxels with voxels in the right and left as neighbouring. Note that this matrix is symmetric, thus voxel couples are counted twice. For the GLRLM, we also used the same definition of the neighbourhood as for the GLCM. For all other matrices, the neighbourhood is given by the infinity norm. The radiomic features presented are only representative examples. 13

4.1 Kaplan Meier plot on the 561 TCGA-GBM patients with survival available. Duration is given in days 39

4.2 Kaplan Meier plot on the 561 TCGA-LGG patients with survival available. Duration is given in days. 41

4.3 The different pre-processing steps applied to MR images 43

4.4 Intensity distribution on the MRI image for each patient on the LGG dataset. Columns correspond to the modalities T1w, T2w, FLAIR, CE T1w. First row represent distribution of the white matter voxels, and second row for the grey matter. (a) Before white-stripe and (b) after white-stripe normalisation 45

5.1 Top row: Example of T2w MRI scan of a patient with DIPG tumour extending beyond the pons. (a) Axial, (b) Coronal and (c) Sagittal slices. Bottom row: Example of glioblastoma and DIPG MRI scans, in axial slices. (d) glioblastoma T2w, (e) glioblastoma FLAIR, (f) DIPG T2w and (g) DIPG FLAIR. Tumours appear on the images as hyper-signal. 51

5.2 The two approaches pYU and sYBBU 53
5.3 Experimental design diagram 57
5.4 Details of the 6 different models that are trained for each modality. These 6 models will be used for inference either independantly, or serialised parallelized. 58
5.5 The trained models are used for inference either independently, or serialised parallelized. Finally results may be ensembled accross the modality. 58
5.6 Mean precision-recall graphs of the different proposed segmentations on the TCGA-GBM dataset. To focus on the most interesting part of the plot, we only plotted the precision-recall scores for thresholds between 0.1 and 0.9. From the left to the right, Using the FLAIR, using the T2w, and ens.(FLAIR,T2w). 61
5.7 Mean precision-recall graphs of the different proposed segmentations on the LGG dataset. To focus on the most interesting region, we only plotted the precision-recall scores for thresholds between 0.1 and 0.9. From the left to the right, using the FLAIR, using the T2w, and ens.(FLAIR, T2w) 62
5.8 FLAIR image of a DIPG patient from the PREBIOMEDE cohort. Detection Failed on the image. 64
5.9 Segmentation results obtained on one case of DIPG are superimposed on the FLAIR background. The top row displays the complete patient images. Ground truth mask in blue. Yolo detection contours in red. sYBBU Segmentation with FLAIR in orange. sYBBU Segmentation with T2w in green. pYU Segmentation with FLAIR in purple. pYU Segmentation with T2w in yellow. The tumours are presented as a hyper-signal. 66

6.1 Star and Path graphs with different u_2 values. Grey nodes correspond to 0 values, red for 1, and blue for -1 77
6.2 Illustration of a case were $\mathbf{\Sigma}_2$ does not reflect a valid correlation matrix 77
6.3 Details of the proposed experimental design for the TCGA-LGG dataset. The analysis step refers to the comparison between the different graphs, analysis of the variable selection when nodes are permuted or edges are removed. Application refers to survival prediction and enrichment analysis. 81
6.4 Comparison between Raw and Normalised Laplacian. (a) Evolution of the number of selected genes as $\gamma_{\mathcal{G}}$ varies, using the raw and normalised Laplacian. (b) Evolution of the stability metric as $\gamma_{\mathcal{G}}$ varies, using the raw and normalised Laplacian. (c) Number of selection co-occurrence of selected genes, with $\gamma_{\mathcal{G}} = 10^2$. 83
6.5 Degree distribution of selected genes by fold for Raw and Normalised Laplacians. (a) For each selected gene, we counted the number of its neighbours in the PC graph. The black line represents the density of the degree distribution of all genes in the PC graph. (b) For each selected gene, we counted the number of its neighbours among selected genes. 83
6.6 Box plot of the weight distribution of selected genes for Raw and Normalised Laplacians. x-axis is log2 of the gene degrees in full PC graph. 84
6.7 Average absolute weight difference between selected nodes according the Dijkstra distance between them in the PC graph 84

6.8 Correlation distribution of selected genes. Between stable selected genes (Intersection of selected genes in the 5 folds), Candidate selected genes (Union of selected genes in the 5 folds) and All the genes in the dataset. 85
6.9 Venn diagram showing the overlap between genes selected by the PC graph and the MSIGDB and the KEGG Graphs. Diagrams on the right shows the gene sets resulting from the candidate selected genes, while left diagrams show the gene sets from the stable selected genes 85
6.10 Box plot for the IOU metric between the genes selected by the PC graph and the permuted PC graph 86
6.11 The evolution of the number of selected genes when the number of edges decreases (x-axis is the percentage of edges removed from 0% to 90%.) 86

7.1 Diagram of the different run models and the different comparisons discussed in this chapter. 95
7.2 Variable selection and model assessment methodology. The same CV folds and bootstrap samples are kept for all compared models. 96
7.3 Distribution of the C-index values according to the modality of the radiomics. Obtained from the ten bootstrap runs 98
7.4 Distribution of the C-index values according to the pre-processing of somatic mutation profiles. Obtained from the ten bootstrap runs. 101
7.5 Number of times the selected genes appear to be mutated in the original somatic mutation profiles, using graph propagation (left) with the PC graph, and (right) with the influence graph. 103
7.6 Distribution of the C-index values obtained from the ten bootstrap runs for RNA. 103
7.7 Distribution of the C-index values obtained from the ten bootstrap runs for RNA. 105
7.8 Number of radiomic features selected from each modality (left) at least in one run (right) in all three runs. 106
7.9 (a) Number of selected variables from each block, in each successful run using the Cox model with all blocks concatenated. (b) Total number of variables from each block using the Cox model with all blocks concatenated 108
7.10 Distribution of the C-index values obtained from the ten bootstrap runs, for the four multiblock models tested. 109
7.11 Correlation matrices between the 17 most frequently selected genes from the RNA block by the netSGCCA model. (left panel): the minimum correlation between the genes in the ten bootstrap samples; (center panel): the maximum correlation between the genes in the ten bootstrap samples; (right panel): the difference between the two previous matrices. . . 110

List of Tables

4.1 TCGA-GBM Patient Demographics 40
4.2 TCGA-LGG Patient Demographics 40
4.3 Number of samples available of the different datasets used 42

5.1 YOLO parameters used for training 55
5.2 Dataset sizes of the different training and testing sets 57
5.3 Detection results on TCGA-GBM with 97 test patients. Results present the mean ± standard deviation 59
5.4 Detection results on LGG with 76 test patients. Results present the mean ± standard deviation. 59
5.5 Segmentation results on TCGA-GBM with 97 test patients. Unlike (BB-)UNet approaches, Deepmedic network (Kamnitsas et al., 2017) is trained on 3D volumes from the HGG^{train} dataset. Results present the mean ± standard deviation. 60
5.6 Correlation study. Correlation values between detection precisions and final segmentation results obtained on the ensembled bounding-boxes. 61
5.7 Correlation study. Correlation values between detection recalls and final segmentation results obtained on the ensembled models. 61
5.8 Comparison of segmentation performances when using the real bounding-boxes and predicted bounding-boxes for the FLAIR without post-processing. Results present the mean ± standard deviation. 62
5.9 Segmentation results on LGG with 76 test patients. Results present the mean ± standard deviation 63
5.10 Detection results on 71 test cases from the DIPG set. 63
5.11 Detection results on 62 sessions from the DIPG dataset excluding the 9 sessions where detection step failed to detect anything. Results present the mean ± standard deviation . 64
5.12 Segmentation results on 62 test sessions from the DIPG set, using ens.(FLAIR, T2w) detection masks. 64
5.13 Segmentation results on 62 sessions from DIPG, using FLAIR detection masks. Results present the mean ± standard deviation 65

6.1 Recovering performances depending on configurations defined by the different cases defined by the vector $\boldsymbol{u}_2$ and graphs. Corr is the correlation between the estimated components. Precision, Recall and F1 correspond to the evaluation of $\boldsymbol{u}_2$ against the computed weights. Bold refers to highest values between netSGCCA and SGCCA. Low mean to variance ratio ($c = 0.5$). 79
6.2 Recovering performances depending on configurations defined by the different cases defined by the vector $\boldsymbol{u}_2$ and graphs. Corr is the correlation between the estimated components. Precision, Recall and F1 correspond to the evaluation of $\boldsymbol{u}_2$ against the computed weights. Bold refers to highest values between netSGCCA and SGCCA. High mean to variance ratio ($c = 2$). 79
6.3 Different sources of prior knowledge graphs 80
6.4 The effect of pruning edges that connect genes selected when using the full PC graph. Candidate and stable genes sets obtained when using the different pruned sub-graphs. . . 86
6.5 Performances in survival prediction, number of selected variables and pathways depending on the type of graph to constrain the model. 87
6.6 Top gene sets from MSigDB C6 collection. In bold, gene sets with an adjusted p-value lower than 0.05. 88

7.1 Survival prediction results (C-index) obtained on the 95 patients from the TCGA-LGG dataset using radiomic features. Validation results were obtained on three-fold cross-validation on 76 patients. Test results were obtained using 10 bootstraps on 19 patients. Elastic-Net Cox was used. 97
7.2 Selected radiomic features extracted from the T2 modality. Out of 10 bootstrap samples, we only kept the seven where the model did not fail. The table only shows radiomic features that were selected more than three times. The means and std were computed using only the samples where the feature was selected. 98
7.3 Selected radiomic features extracted from the FLAIR modality. The table only shows radiomic features that were selected more than 6 times. The means and std were computed using only the samples where the feature was selected. 99
7.4 Selected radiomic features extracted from the CE T1 modality. The table only shows radiomic features that were selected more than 3 times. The means and standard deviation were computed using only the samples where the feature was selected. 99
7.5 Selected radiomic features extracted from all modalities. Out of 10 bootstrap samples, we only kept the eight with a convergent model. The table only shows radiomic features that were selected more than 3 times. The means and std were computed using only the samples where the feature was selected. 100
7.6 Survival prediction results obtained on the 419 patients from the TCGA-LGG dataset using mutation features. Validation results were obtained on three-fold cross-validation. Test results were obtained using 10 bootstraps. Elastic-Net Cox was used 101
7.7 Selected genes using the mutation matrix propagated in the PC graph. The table shows genes that were selected more than 5 times. The means and std were computed using only the samples where the feature was selected. 102

7.8 Selected genes using the mutation matrix propagated in the influence graph. The table shows genes that were selected more than 5 times. The means and std were computed using only the samples where the feature was selected. 102
7.9 Enrichr Enrichment results when using the five most selected genes using the influence graph. 102
7.10 Survival prediction results obtained on the 419 patients from the TCGA-LGG dataset using RNA features. Validation results were obtained on three-fold cross-validation. Test results were obtained using 10 bootstraps. Elastic-Net Cox was used. 104
7.11 Sample size comparison between the baseline setting and the current setting devoted to mono- *multi*- block investigation. 104
7.12 Mono-block results obtained on the 83 subjects having the radiomics, somatic mutations and RNA data available. For the Radiomics, the concatenation of all the radiomics extracted from the four modalities has been used; The somatic mutations with graph propagation with the influence graph was used; Finally, the concatenation of all the blocks was also investigated 104
7.13 Selected radiomic features. The table shows features selected in convergent Cox models. The means and standard deviation were computed using only the samples where the feature was selected. 106
7.14 Selected genes using the somatic mutation profiles. The table shows genes that were selected in all 10 runs. The means and std were computed using only the samples where the feature was selected. 107
7.15 Enrichment results when using the ten most selected genes using the influence graph. . . 107
7.16 Enrichment results when using the matrix with concatenated variables from all data sources.108
7.17 Multi-block results obtained on 83 subjects having the radiomics, somatic and RNA data available. For the radiomics, the concatenation of all the radiomics extracted from the four modalities has been used; The somatic mutations with graph propagation with the influence graph was used. 109
7.18 Selected features using the netSGCCA, the normalised Pathway Commons Laplacian was applied on RNA data. The table shows features that were selected more than 5 times. The means and std were computed using only the samples where the feature was selected. . . 110
7.19 Enrichment results when using the most selected genes from the RNA block using netSGCCA. 111

A.1 Selected mutated genes using the raw mutation matrix. Out of 10 bootstrap samples, we only kept the 6 where the model did not fail. Table only shows genes that were selected on all the runs. 116
A.2 Selected genes using the RNA data. Table shows genes that were selected in all runs. . . . 117

List of Algorithms

1 RGCCA optimisation algorithm 71
2 netSGCCA optimisation algorithm 75
3 Priority-Lasso algorithm 94

Chapter 1

Introduction

Gliomas in the WHO classification. Central Nervous System (CNS) tumours are neoplasms located in the brain or spinal cord tissue. Among them, gliomas have the distinctive feature of developing from a glial cell (astrocyte, oligodendrocyte). Glial cells support and protect nerve tissue by providing nutrients and oxygen to neurons; they produce the myelin sheath that makes the transmission of the electric signal or nerve impulse efficient. Gliomas constitute approximately 30% of all CNS tumours and 80% of all malignant brain tumours (Goodenberger and Jenkins, 2012). Until the 2007 edition of the World Health Organisation (WHO), CNS classification for operable tumours mainly relied on the histological comparison between the tumoural cells and their putative original cells (Louis et al., 2016, 2021). The WHO further categorised CNS tumours by their grade, reflecting the tumour aggressiveness, thus the patient's clinical evolution.

According to the 2007 WHO CNS classification of tumours, gliomas included Low-Grade Gliomas (LGG) and Glioblastomas. Low-Grade Gliomas comprised diffuse low-grade and intermediate-grade gliomas, graded II and III by the WHO. These tumours generally impact young, otherwise healthy patients and have an indolent course. LGGs have a longer survival rate in comparison with higher-grade gliomas. The lower grade gliomas are characterised by the IDH mutant and 1p/19q co-deletion (Forst et al., 2014; TCGA, 2015; Wong et al., 2022). On the other hand, Glioblastomas are classified as grade IV. Newly diagnosed patients with Glioblastoma have a median survival time of one year, with poor responses to treatments. This tumour is characterised by alterations in the pathways p53, Rb, receptor tyrosine kinases (RTK) and phosphoinositide 3-kinase (PI3K) signalling, among others (Brennan et al., 2013; McLendon et al., 2008).

Since the early 2010s, several research groups reported differences among the Grade IV gliomas, whether they affected adults (adult High-Grade Gliomas - aHGG) or children (pediatric High-Grade Glioma - pHGG). For example, the diagnostic from histopathology was poorly reproducible in pHGGs because they are heterogeneous (Gilles et al., 2008), and the therapeutic answer was very different for a supposed identical sub-type. The location of some specific pHGG in the brain is more often thalamic and infra-tentorial tumours which distinguished them from the aHGG (Puget et al., 2012).

The 2016 edition of the WHO CNS classification introduced a paradigm shift in the diagnosis of CNS neoplasms. After the recommendations of the consortium cIMPACT which called for a better consideration of the molecular characteristics currently available, both histologic features and genetic alterations were incorporated into the diagnostic framework, classifying and grading brain tumours

(Louis et al., 2021). The WHO CNS classification has had a very recent update in 2021. Under the current version, adult and paediatric HGGs are divided based on an increasing emphasis on molecular markers (Gaillard and Yap, 2017). More specifically, four general groups of diffuse gliomas are recognised in the 2021 WHO classification: 1) adult-type diffuse gliomas, 2) paediatric-type diffuse low-grade gliomas, 3) pediatric-type diffuse high-grade gliomas, and 4) circumscribed astrocytic gliomas (Osborn et al., 2022).

Pediatric Diffuse High-Grade Gliomas. Diffuse Intrinsic Pontine Glioma (DIPG) is a rare brain tumour located in the pons, mostly found in children between 5 and 7 years of age. It is considered one of the most aggressive paediatric tumours, with a survival rate of less than 10% beyond two years after diagnosis (Fisher et al., 2000) and a median overall survival below one year (Cohen et al., 2017). The DIPG is categorised as a diffuse midline glioma (DMG), characterised mainly by a K27M mutation in genes coding for the histone H3 protein and/or a loss of H3K27 trimethylation through EZHIP protein over-expression (Castel et al., 2015). The location of the tumour and its corresponding genomic alteration make the DIPG a particular type of tumour from other High-Grade Gliomas (HGGs) (Louis et al., 2016).

DIPG grows fast. Its symptoms include eye problems, trouble with walking, muscle coordination or balance, and weakness in the arms or legs. Diagnosis of DIPG principally relies on Magnetic Resonance (MR) Images, but they are insufficient as the tumour can be confused with other similar tumours. Biopsies can be safely used for diagnosis (Gupta et al., 2018). However, it is rarely done and has yet to be incorporated into the standard diagnosis protocol.

Currently, there is no curative treatment for patients with DIPG. Removing the tumour with surgery is not feasible due to its location. The tumour is in the brainstem, which helps control essential functions. The tumour can not be removed without damaging vital brain tissue. Most patients are treated with radiation therapy, which has been shown to increase survival and improve symptoms temporarily in most patients (Gallitto et al., 2019). Chemotherapy has been found challenging due to the blood-brain barrier. The blood–brain barrier is a highly selective semipermeable border that prevents solutes in the circulating blood from non-selectively crossing into the extracellular fluid of the central nervous system. However, some chemotherapy treatments are currently being examined in clinical trials (Gwak and Park, 2017).

Multiple clinical trials have been performed to investigate DIPG therapies (Rechberger et al., 2020). Some trials focused on progression-free survival or overall survival (Burzynski et al., 2014; Kebudi et al., 2019). While some results are promising, the overall landscape is grim. Very few results have been translated into practice as of yet. This makes it critical to use new strategies and novel technologies to uncover new biological factors relevant to tumour development and its cure.

Magnetic Resonance screenings have been incorporated into most neuro-oncological diagnosis protocols and are well established in WHO CNS for non-invasive tumour characterisation. They have become prominent among various medical imaging techniques due to their safety and information abundance. These images are routinely used for pathology management, including detection, biopsy guidance, and treatment response evaluation.

Recent years have known multiple advances in medical imaging technologies, especially MR imaging. These advances include improvements in image quality and spatial resolution. This can be

attributed to higher magnetic field strengths, using 3T magnets instead of 0.5T and 1.5T several years ago. The advances are not limited to hardware. Developments in software have led to increased imaging speeds, which reduced motion artefacts and screening costs, thus making MRI screening more available, accessible and affordable. Further details about MR images will be discussed in the Chapter 2.

DIPG and radiomics. In the case of WHO CNS subtype 3 diffuse high-grade gliomas, patients now undergo MRI exams for diagnosis but also for the follow-up of the answer to treatment. Recently (Aerts et al., 2014) proposed to systematically extract textures of the computed tomography (CT) image in the different compartments of lung and head-and-neck tumours to study their potential prognostic power. This approach, which was generalised to other imaging types, including MRIs and was coined radiomics, allows the systematic exploration of the tumour macro/micro organisation depending on the imaging sequence and resolution at hand. Practically, radiomic analysis includes all the extraction and analysis steps of a large number of features (200+) from a region of interest (ROI) on the available MRIs. These features include size and shape characteristics which describe apparent visual properties, then first-order statistics of the signal from the ROIs. Additionally, radiomic features include second-order statistics (texture) features that describe fine, local grey-level configurations. The hypothesis behind radiomic analysis is that MRI can capture texture features invisible to the naked eye, which can be linked to a disease outcome.

DIPG and high-throughput genomics. When a biopsy can be performed, the diagnosis of DIPG includes the use of advanced molecular immunohistochemistry tools. But for research purpose, high-throughput genomics is now often considered. Indeed in parallel to imaging progress, recent years have also known breakthroughs in molecular data acquisition technologies such as DNA/RNA Sequencing and Single Cell Sequencing, which increased the accessibility of these technologies. Coupled with the decrease in data storage costs and the rise in computational capabilities, these have made high-throughput multi-modal databases for hundreds of patients publicly available to study clinically relevant problems, especially in oncology.

DIPG and multi-view measurements. The recent evolution of data available in clinical research for DIPG result in composite data obtained with various measurement tools from the same observations. It includes, but is not limited to, multi-omics data such as gene expression profiles, mutation profiles, copy number variants, clinical data from standardised electronic Case Report Forms, and imaging data. These data are regrouped into separate views or blocks, each block including multiple variables. Analysing these datasets required the development of new multi-view machine-learning and statistical models designed to integrate data (Cantini et al., 2021; Herrmann et al., 2020). Unlike traditional machine-learning models that account for each block contribution separately, multi-block models allow taking into account the information shared among blocks and the interactions between variables across the blocks (Philippe, 2014).

The methodological work presented here is devoted to subtype 3 in WHO classification of CNS tumours, which we will name DIPG in this manuscript from now on. It aims at providing analysis tools of molecular and biomedical imaging profiles to inform treatment decisions to control the disease.

Because of data sparse availability, our developments will leverage measurements in other subtypes when the category of measurements is similar across subtypes.

1.1 Open questions raised by multi-view datasets in rare tumours

Data presentation. DIPG or DMG are rare tumours with median overall survival of less than one year. This situation motivates clinical research projects aiming to acquire rich and multiple molecular measurement datasets obtained at different levels of biological organisation. These datasets could open new therapeutic opportunities and be used to derive different disease subtypes. In DIPG, tumours are inoperable because they are highly infiltrative and located in the pons. Consequently, the disease evolution and treatment impact on the tumour (radiotherapy and possibly targeted chemotherapy) is monitored using MR imaging assessment.

Most of these clinical projects are pharmaceutical – they intend to study promising protocols - and include ancillary data for further study. Two clinical trials (PHRC BIOMEDE [1] and PHRC BIOMEDE 2.0 [2]) have been recently designed and promoted by Gustave Roussy Cancer Campus. They aim to inform treatment decisions according to molecular biomarkers and test targeted chemotherapies assigned based on patients' individual molecular profiles. This is to improve disease control along with conventional radiotherapy. The BIOMEDE pharmacological trials also include an ancillary study that brought additional data consisting of multi-omics from the biopsy, imaging acquired longitudinally, image biomarkers resulting from image processing, and clinical data. Even if the data of the ancillary study are not yet wholly available, they foreshadow the context of this type of study with multi-block data. In addition to their multi-block presentation, they include a significant amount of missing data. The most salient point about these data is the small number of samples available, which, no matter how much effort is made, will remain limited given that this is a rare disease.

Data about DIPG is scarce, and knowledge about the disease is limited. This situation can, however, be mitigated by the availability of a large amount of public data. These are of two kinds. Firstly, there are data on other gliomas that are similar in some respects to DIPGs and contain the same types of measurements collected for BIOMEDE. These include data from the iconic *The Cancer Genome Atlas (TCGA)* [3] and *The Cancer Imaging Archive (TCIA)*[4], projects which offer molecular and imaging data on many patients. These public datasets have the advantage of having a large number of available samples and extensive literature about them. Therefore, they can be used to study the behaviour and limitations of the methods used, and the results can be compared with existing literature. Additionally, while we know that DIPG is a unique tumour type, there exists, nevertheless, similarities with other gliomas that can be exploited to learn about the disease. The second kind of publicly available data is the curated and extensive knowledge about molecular data. One can cite the *Molecular Signature Database (MSigDB)*[5] of the Broad Institute, which compiles collections of cancer signatures in the form of gene lists, and the *PathwayCommons* [6], which compiles

[1] https://clinicaltrials.gov/ct2/show/NCT02233049
[2] https://clinicaltrials.gov/ct2/show/NCT05476939
[3] https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga
[4] https://www.cancerimagingarchive.net/
[5] https://www.gsea-msigdb.org/gsea/msigdb
[6] https://www.pathwaycommons.org/

all the interactions described between different molecular entities in the form of interaction graphs. These data can be used as prior knowledge of our methods and help the interpretation of the results.

It should also be noted that data scarcity is sometimes due to shortcomings in pre-processing methods adapted to data acquired under routine clinical conditions. The systematic and rational extraction of standard information from MR images requires numerous processing steps, such as the co-registration of images acquired with different sequences, their normalisation, harmonisation and quantification. Most of these steps have been implemented in clinical neuroscience work that operated on images without mechanical deformation, such as that caused by tumours. It is, therefore, necessary to revisit these approaches, and we have done this work on the data considered in our work. In contrast to these methods, whose adjustment does not constitute original work, the automatic delineation of DIPG/DMG type tumours constitutes an open research challenge. Indeed, the pHGG tumour cases are absent from the data compendiums used to train the models for automatic tumour delineation.

Overall, the presentation of the DIPG **calls for methods to fully potentiate the images to generate the radiomics** information we wish to study. This presentation of data also calls for models capable of integrating the different types of data and training paradigms that allow to take advantage of external information: these paradigms can consist in transfer learning - where a model trained on external data is only refined - or consist in penalising (or orienting) existing statistical models.

Non-black-box models are desired. In the data setting described above (collected from clinical trials combined with public data), statistical models are not only used to predict the outcome of diseases. Instead, they are expected to help to characterise the disease by finding a set of variables of interest explaining their outcome.

However, isolated variables associated with a pathology outcome usually do not, by themselves, give a biological meaning to this association. For example, at the gene scale, somatic mutation profiles from patients with the same pathology may differ from one to the other because different genes are mutated in the same pathway (Lawrence et al., 2014; Le Morvan et al., 2017; Wood et al., 2007). This leads to the need to identify networks and pathways that group variables that interact in complex patterns. Several works have proposed the usage of complex graphs to identify sub-networks of variables of interest (Zhang et al., 2017). Some works suggested using graphs as a post-analysis tool to identify the interactions between previously selected variables (Kim et al., 2011; Vandin et al., 2011a; Vaske et al., 2010). Others used the graphs as a pre-processing tool, such as smoothing variables over the interaction network (Hofree et al., 2013; Le Morvan et al., 2017; Rapaport et al., 2007). Graphs have also been integrated into statistical models as a penalty over the model parameters. This includes supervised models, such as survival models (Zhang et al., 2013) and regression models (Li and Li, 2008), and unsupervised models, such as matrix factorisation models (Zhu et al., 2021).

Overall, the expectations regarding statistical models used in clinical research projects concern the interpretability of these models and their ability to propose molecular hypotheses for treatment pathways or operating modes. This implies, for example, providing existing multi-block approaches with penalties so that the models selected by the learning procedure offer interpretable solutions: for example, we can try to constrain that the set of gene expression variables selected are "close" in the PathwayCommons graph and therefore belong to a same biological pathway. To **contribute to the**

expected interpretability properties, graphical penalisation is scheduled on at least one block of the multi-block models. This penalty on models must scale to accommodate the block with a large number of variables, and it is necessary to have a framework to evaluate their contribution. To provide interpretations even if few clinical annotations are available in a cancer sample, it is necessary to consider statistical models that are more than mere classifiers but also survival models. **Indeed, censored survival data is generally available. Yet it implies adapting the training framework of statistical models**.

Finally, the question of the unitary and joint contribution of the different blocks arises in the heavy and complex experimental setting of collecting molecular and imaging data. **It is necessary to evaluate these contributions under conditions of data scarcity**. In addition, the question arises of the transfer learning approaches that could be constructed to move from one imaging + multi-omics dataset to another.

1.2 Manuscript outline

This work aims at studying the integration of radiomics with omics data in a multi-block analysis framework. Many issues are raised by the multi-block integration in DIPG. We addressed a few of them, highlighted in the previous paragraph. To achieve these tasks, we do not only consider the specific DIPG dataset (as it will be made available by the BIOMEDE clinical research). Instead, we consider several datasets of interest, namely LGG, Glioblastoma and DIPG. In some tasks, we used explicitly transfer learning, but in other, we implemented and evaluated methods to anticipate complete data annotation availability. Since radiomics are extracted from regions of interest in the images, in our case, the tumoural area, this work discusses brain tumour segmentation approaches.

The first part of this manuscript presents the basic definitions used and the related works. Then, it introduces the different datasets and the various preprocessing methods applied.

The second part is organised into three chapters. Chapter 5 presents our brain tumour segmentation approach, already published (Chegraoui et al., 2021a) that is demonstrated on DIPG data. In Chapter 6, we lay out our multi-block integration approach penalised with a graph *a priori*. The method is assessed on simulations and TCGA-LGG (TCGA, 2015) data. Finally, in Chapter 7, we study the potential of adding imaging to molecular data by applying our methods on the TCGA-LGG dataset, including imaging and molecular blocks. We show results obtained on the survival prediction with downstream analysis.

* * *
* *
*

Part I

Materials and Methods

Chapter 2

Notations and Definitions

In this chapter, the different notations and definitions of entities discussed in the thesis are introduced. This chapter also includes the definition of classical methods discussed in this thesis.

Chapter Outline

Contents

2.1 Notations 10
2.2 Data description 10
2.2.1 Imaging Data 10
2.2.2 Genetic data 14
2.3 Method Definitions 14
2.3.1 Objective functions 15
2.3.2 Regularisation 15
2.3.3 Classical models 16
2.3.4 Metrics 19

2.1 Notations

Throughout this work, we use the following convention. Bold uppercase letters represent matrices, such as $\boldsymbol{A}$. Bold lowercase represents vectors, such as $\boldsymbol{a}$. Scalars are represented by lowercase letters, such as a. We denote with a star an optimal parameter (or vector) as $a^\star$. Data is represented in matrices, often denoted $\boldsymbol{X}$, with n observations (rows) and p variables (columns). These data matrices are also called blocks. In this work, we are interested in the analysis of J blocks, which we denote $\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(J)}$ and their associated vectors are denoted $\boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(J)}$ ($\boldsymbol{\beta}$ can also be used in some cases).

We define a graph as $\mathcal{G} = (\mathcal{E}, \mathcal{V})$, where $\mathcal{E}$ is a finite set of nodes and $\mathcal{V}$ the set of edges. Unless specified otherwise, the graph is undirected. For sake of simplicity, the elements $e_1, e_2, \cdots$ of $\mathcal{E}$ can be referred to as $1, 2, \cdots$. The order of the elements is arbitrary but consistent. To the graph $\mathcal{G}$, a binary adjacency matrix $\boldsymbol{A}$ is associated, where $\boldsymbol{A}_{i,j} = 1$ if the nodes i and j are connected, otherwise $\boldsymbol{A}_{i,j} = 0$. Each element of the graph has a degree counting its number of neighbours, which can be represented in a diagonal matrix $\boldsymbol{D}$. The laplacian $\boldsymbol{L}$ associated with the graph $\mathcal{G}$ is defined as $\boldsymbol{L} = \boldsymbol{A} - \boldsymbol{D}$.

We denote the norm operator $\|\cdot\|$. The ℓ_1 norm of a vector $\boldsymbol{\beta}$ is defined as $\|\boldsymbol{\beta}\|_1 = \sum |\boldsymbol{\beta}_i|$. The norm ℓ_2 is defined as $\|\boldsymbol{\beta}\|_2 = \sqrt{\sum \boldsymbol{\beta}_i^2}$. Let $\boldsymbol{K}$ positive-defined matrix, the norm $\|\boldsymbol{\beta}\|_{\boldsymbol{K}} = \sqrt{\boldsymbol{\beta}^\top \boldsymbol{K} \boldsymbol{\beta}}$. For a real matrix $\boldsymbol{M}$ we define the Frobenius norm noted $\|\boldsymbol{M}\|_F = \sqrt{trace(\boldsymbol{M}^\top \boldsymbol{M})}$. Recall that the trace function returns the sum of diagonal entries of a square matrix.

2.2 Data description

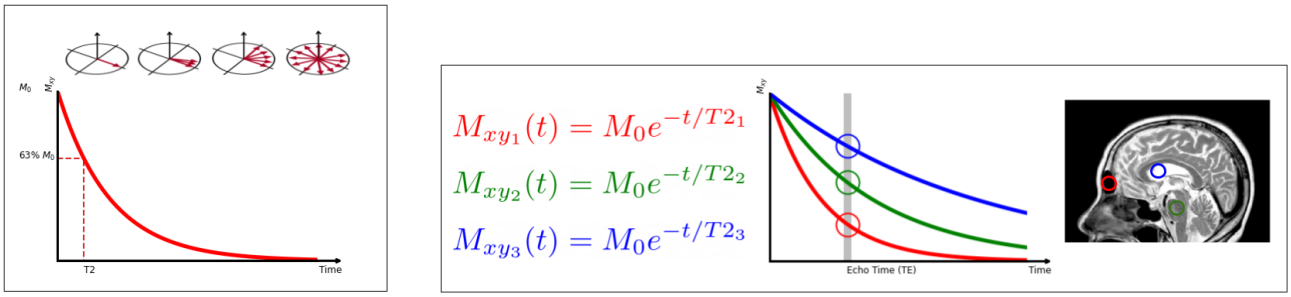
In this thesis, we are interested in two sources of data. First, imaging data was acquired through MR scans and described using radiomic features. Second, biological samples were obtained through biopsy, from which genetic data could be extracted. Here, we are mainly interested in gene expression, somatic mutation profiles and copy number variants (CNV).

The following sections will describe the different data used, starting with imaging and then genetic data.

2.2.1 Imaging Data

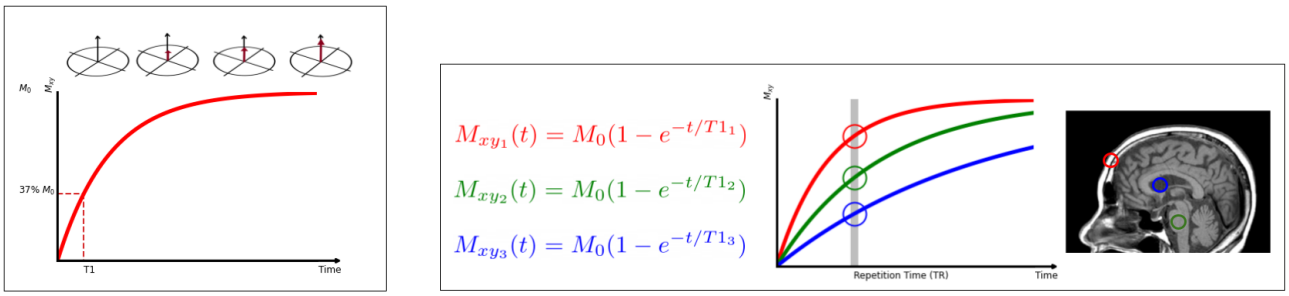
Images are acquired through an MRI scanner composed of a magnet, gradient coils, and radio-frequency coil. The magnet generates a magnetic field with strength in the order of a few teslas. The gradient coils are responsible for spatially localising the signals. Finally, the radio-frequency coils generate and send radio-frequency signals in the order of a few microteslas. The scanner exploits the intrinsic spin property of protons to generate the images. MRIs mainly focus on protons in hydrogen in the water molecules. First, the magnetic field aligns the spin of the protons with its direction. Radio frequencies are sent towards the protons, via the radio-frequency coil, which knocks them out of alignment. As the spin realigns with the magnetic field, they generate an electrical signal which fades through time due to the desynchronisation of spins. The decay of this electrical signal is characterised by its T2 relaxation time and is tissue specific. We can then measure the contrast between the decay-

ing electrical signals at an operator-chosen Echo Time (TE). In parallel, as the protons realign with the magnetic field, their magnetism increases. This growth rate is tissue-specific and characterised by a T1 relaxation time. We can then measure the contrast between the increasing magnetism at an operator-chosen Repetition Time (TR). It should be noted that these two signals cannot be measured simultaneously. Figure 2.1 and Figure 2.2 illustrate the acquisition of the described signals.



((a)) ((b))

Figure 2.1: Acquired T2 signal. (a) As the spins realign with the magnetic field, their electrical signal decreases. The decrease is characterised by the T2 relaxation time. (b) The T2 relaxation time is tissue dependant. Acquiring the signal at TE gives the contrast between the different tissues and results in a T2-weighted image.



((a)) ((b))

Figure 2.2: Acquired T1 signal. (a) As the spins realign with the magnetic field, their magnetism increases. The increase is characterised by the T1 relaxation time. (b) The T1 relaxation time is tissue dependant. Acquiring the signal at TR gives the contrast between the different tissues and results in a T1-weighted image.

Choosing TE and TR results in different MR sequences (or modalities), which capture various characteristics of the tissue organisation (at a millimetric or submillimetric scale) or functional parameters. These sequences differ by how they contrast the different tissues or map functions in tissues. Four modalities of MR images are commonly used for clinical diagnosis: The T1-weighted (T1w), the T2-weighted (T2w), Fluid Attenuated Inversion Recovery (FLAIR) and Contrast Enhanced T1-weighted (CE T1w). The T1-weighted images are obtained with a short TE and a TR similar to the T1 relaxation time. The T1-weighted modality highlights fatty tissues and is best for observing anatomical structures such as the grey and white matter in the brain. CE T1w scans are T1w scans after the patient's infusion of a contrasting agent, usually, Gadolinium (Gad), a non-toxic paramagnetic agent. CE T1w images are especially useful in looking at vascular structures and breakdowns in the blood-brain barrier. The T2-weighted images are acquired with a TE similar to the T2 relaxation time

and a long TR. T2-weighted images are bright on the fat tissues and fluids. The FLAIR acquisition parameters are similar to the T2-weighted. However, additional pulses are sent to the protons in the acquisition phase to nullify the signal from fluids. The contrast in the FLAIR images is similar to the T2-weighted, but the intensity of normal fluids is attenuated. Both T2-weighted and FLAIR sequences are utilised for tumour location.

2.2.1.1 Radiomics

Radiomics is defined as the extraction of several feature descriptors from ROIs in a discretised image. Various features can be extracted from the image signal and they are generally grouped into several classes. Shape-based features account for 26 descriptors, including the ROI volume, surface area and sphericity. First-order features comprise 19 metrics describing voxel intensities within the ROI. These metrics include energy, entropy and the mean of intensities. Finally, the texture features are metrics that summarise the content of five co-occurrence matrices: the Gray Level Co-occurrence Matrix (GLCM), the Gray Level Run Length Matrix (GLRLM), the Gray Level Size Zone Matrix (GLSZM), the Neighbouring Gray Tone Difference Matrix (NGTDM) and the Gray Level Dependence Matrix (GLDM).

The GLCM counts the number of times two intensities are neighbouring each other. Formally, let $\boldsymbol{M}$ be the GLCM, $\boldsymbol{M}_{i,j}$ is the number of times a voxel of intensity i is at maximum a distance δ from a voxel of intensity j. This matrix is normalised to reflect the joint probability distribution of image intensities. From the GLCM, 24 features are extracted, including Autocorrelation, Contrast and Joint Average.

The GLRLM counts the number of same intensity level sequences by length. Formally, let M be the GLRLM, $\boldsymbol{M}_{i,j}$ is the number of sequences of intensity level i and length j occurring in the ROI. From the GLRLM, 16 features are extracted, including Short Run Emphasis, Long Run Emphasis and Gray Level Non-Uniformity.

The GLSZM quantifies connex grey level zones with the same intensity level. Let $\boldsymbol{M}$ be the GLSZM, $\boldsymbol{M}_{i,j}$ is the number of connex components of size j of intensity level i. Two voxels of the same intensity are neighbouring if they are adjacent in the image. From the GLSZM, 16 features are extracted, including Small Area Emphasis, Large Area Emphasis and Gray Level Variance.

The GLDM quantifies grey-level similarity in an image. For a parameter δ, two voxels are similar if their intensity levels g_i and g_j verify $|g_i - g_j| \leq \delta$. Formally, let $\boldsymbol{M}$ be the GLDM, $\boldsymbol{M}_{i,j}$ is the number of times a voxel of intensity i has j similar voxels at a distance δ. From the GLDM, 14 features are extracted.

The NGTDM quantifies the difference between an intensity level value and the average intensity value of its neighbours. For each voxel of intensity level i, we compute the absolute value of the difference between i and the average intensity level of its adjacent voxels. Then, these differences are summed across all voxels of the same intensity level (s_i). Additionally, the normalised score $p_i = \frac{s_i}{\Sigma s_j}$ is computed for each intensity level. The two scores are used to calculate 5 radiomic features.

An illustration of the different texture matrices is given in figure 2.3. Multiple software can be used to extract the radiomic features such as LifeX (Nioche et al., 2018) and pyRadiomics (van Griethuysen et al., 2017).

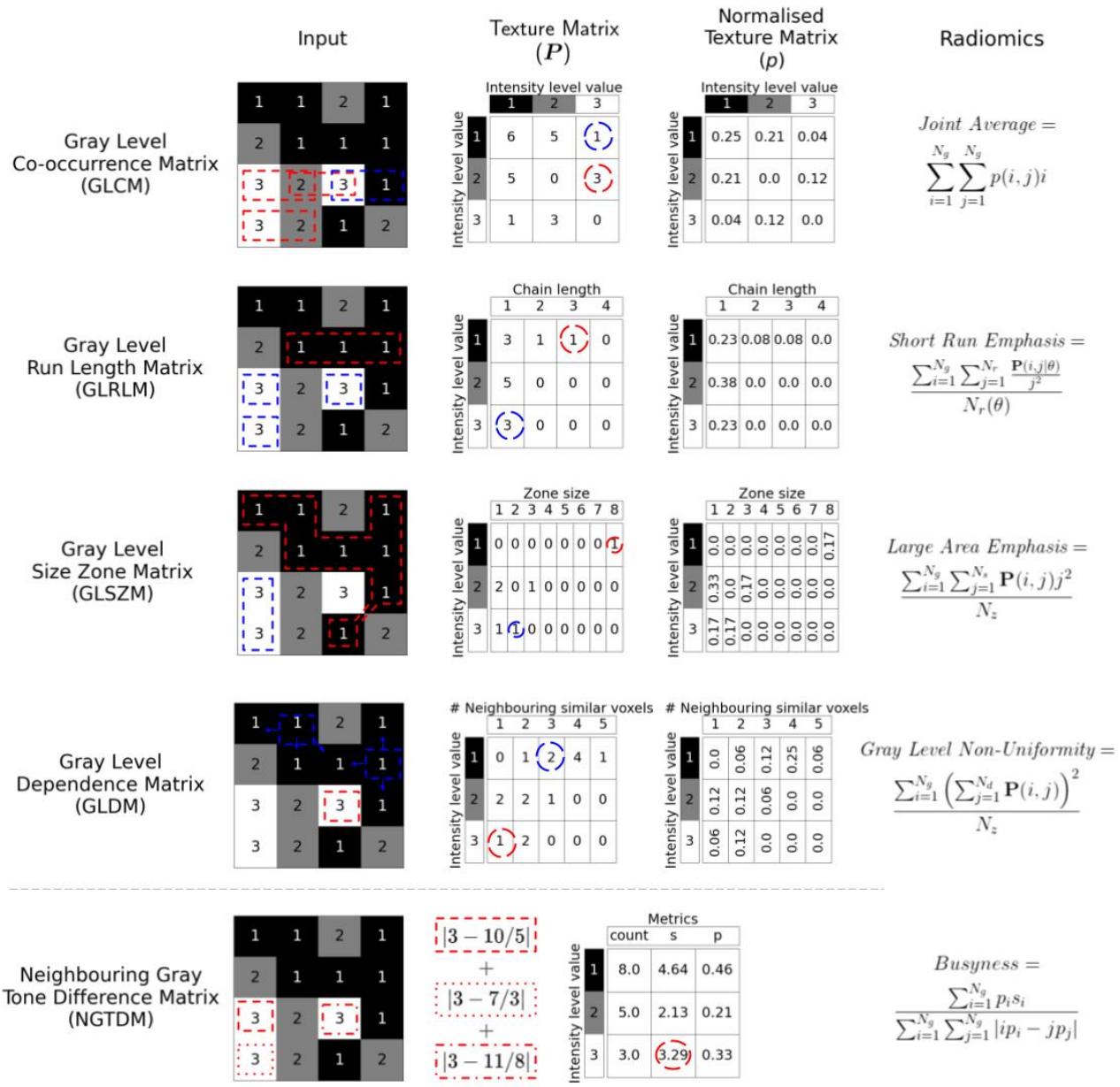


Figure 2.3: A description of how the different texture features are computed. In an example 4×4 image ROI, three gray levels are represented by numerical values from 1 to 3. The red and blue annotations are used to highlight the different voxels used in order to compute some matrix values. To illustrate the different matrices: For the GLCM we only considered voxels with voxels in the right and left as neighbouring. Note that this matrix is symmetric, thus voxel couples are counted twice. For the GLRLM, we also used the same definition of the neighbourhood as for the GLCM. For all other matrices, the neighbourhood is given by the infinity norm. The radiomic features presented are only representative examples.

2.2.2 Genetic data

A biological sample from the patient can be extracted through a biopsy. This sample can then be analysed to obtain different generic profiles.

2.2.2.1 RNA sequencing

Gene expression measures the relative abundance of messenger RNA (mRNA) in a sample prepared from the tumour cells. This technique measures the expression level of a gene transcribed into an mRNA (in the form of one or several splicing variants for a given gene). Raw read counts correspond to the number of times an mRNA sequence was seen in the biological sample. To measure gene expression, RNA-seq is usually used. First, the mRNAs are isolated and then converted to a more stable double-stranded DNA and amplified. A sequencer reads the sequence from each fragment, which is then mapped to a gene and counted.

The read counts are said to be relative since they cannot be used to compare expression levels between samples due to the need to account for differences in transcript length and the total number of reads per sample. Multiple RNA-seq normalisations have been proposed. In this thesis, we used the Transcripts Per Million (TPM) normalisation. For a sample i and a gene j, the TPM matrix is defined as follows :

$$TPM_{i,j} = \frac{q_{i,j}/l_j}{\sum_k q_{i,k}/l_k} 10^6 \tag{2.1}$$

where $q_{i,j}$ is the raw read count for the gene j of the sample i, and l_j is the transcript length of gene j.

2.2.2.2 Somatic mutations

Somatic mutations are mutations observed in somatic cells only, as opposed to germline cells. These mutations can not be passed on to descendants and occur after conception. Mutation profiles do not allow for the identification of the subjects because they are obtained by subtracting a patient's genome from the genome observed in the tumour cells. In this thesis, somatic mutation profiles are represented as a matrix of n subjects and p genes where $\boldsymbol{X}_{i,j} = 1$ if a mutation occurred in the gene j for subject i.

2.2.2.3 Copy number variation (CNV)

Copy number variation (CNV) is the estimation of the variation in the copy number of genomic segments. The CNV is given by the ratio of the estimated concentration of the target gene to the estimated concentration of the reference gene. Sequencing can be used to detect genetic variants, which include CNVs.

2.3 Method Definitions

Machine learning (ML) is devoted to choosing and building mathematical models whose parameters are adjusted according to sets of available data. We define a few general terms used in ML that we

will utilise in our work. Most statistical models define an objective function $f(\boldsymbol{X};\boldsymbol{\beta})$, where $\boldsymbol{X}$ are the observed data, and β are the model parameters. The aim is to find the optimal set of parameters that minimises (or maximises) f. Machine learning approaches are traditionally divided into two categories. The supervised models are used when a desired outcome to predict is available. The model learns a function g that maps the input $\boldsymbol{X}$ to the output $\boldsymbol{Y}$. The unsupervised models have no labels given to the learning algorithm, leaving it on its own to find structure in its input.

This section defines several classical objective functions and models which are discussed in our work. We also present regularisation techniques that add constraints to the models. These constraints help the models discard unwanted noise and prevent it from closely fitting the observed samples without the ability to generalise on new observations. Finally, we discuss some metrics used to assess the model performances, allowing us to compare different models objectively.

2.3.1 Objective functions

Over the years, multiple objective functions (also called loss functions) have been defined. In this section, we describe the most commonly used objective functions.

2.3.1.1 Least-Square

The Least-Square objective function aims to find the best parameters for a mapping $g(\boldsymbol{X};\beta)$, which will minimise the sum of squared errors between the observed data and the predicted data. The observed data come here as a couple $(\boldsymbol{X};\boldsymbol{Y})$. This loss function is formulated as follows:

$$f((\boldsymbol{X};\boldsymbol{Y}),\beta) = \|\boldsymbol{Y} - g(\boldsymbol{X};\beta)\|_2^2 \tag{2.2}$$

Note that the least-square is only for supervised models. g can be a generative model that learns the best models that generate observed data $\boldsymbol{X}$ (In this case, the input $\boldsymbol{X}$ is also the approximated output). This loss function is generally used for regression or image denoising problems.

2.3.1.2 Log-likelihood

The likelihood $f(\boldsymbol{X};\beta)$ is the probability of the realisation of the observed data $\boldsymbol{X}$ considered as a function of β. If the observed samples are assumed to be independent, then the likelihood function is the product of the likelihood of each sample. The log-likelihood is often used to transform the product into a sum. The statistical model aims at maximising the log-likelihood. The log-likelihood is often used for classification and survival models.

2.3.2 Regularisation

Three main possible benefits can arise from adding regularisation over the parameters of a statistical model. First, it can be added to ill-defined objective functions. For example, the loss function f does not have a defined local or global minimum. Adding regularisation limits the search space and thus guarantees the existence of a solution. Additionally, regularisation can be used to prevent the statistical model from over-fitting the data, thus capturing unwanted noise and failing to generalise

on new observations. Finally, regularisation helps retain a subset of the predictors, which helps interpretation.

In order to restrict the search space for the parameters or impose regularity over them, penalties can be added to statistical models. Three main penalties are widely used. First, the Ridge penalty Hoerl and Kennard (1970), which introduces an ℓ_2 penalty over the model parameters, which helps shrink them. This penalty helps solve the multi-colinearity in the data and improves parameter estimation. The Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996) is another popular machine learning regularisation technique. This penalty introduces an ℓ_1 norm over the model parameters to promote sparsity, enhancing the model interpretability. The LASSO selects a few representatives among highly correlated variables in the observed data. Finally, ElasticNet (Zou and Hastie, 2005) is the linear combination of the Ridge and LASSO regularisations. The Elastic-Net performs variable selection while benefiting from the smoothness of the estimates introduced by the Ridge penalty.

2.3.3 Classical models

2.3.3.1 Supervised models

Having observed data or covariates $\boldsymbol{X}$ and target data $\boldsymbol{Y}$, a supervised model finds the optimal parameters that map the covariates to the target. In this section, we present some commonly used supervised models.

Support Vector Machines Support vector machines (SVM) (Cortes and Vapnik, 1995) are classical ML methods. Given a set of labelled samples divided into two categories, an SVM training algorithm builds a model that assigns new samples to one category or the other by finding an optimal hyperplane that separates the two classes. The hyperplane defines a decision boundary between the different classes. Kernels can be used to learn high-dimension feature space. The SVM were initially introduced for two-classes classification tasks, but they can be extended to regression and survival tasks.

Partial Least Squares Partial Least Squares (PLS) (Kramer, 1998) is a linear regression model that maps the input $\boldsymbol{X}$ and target data $\boldsymbol{Y}$ by projecting both matrices into a new space. The underlying model is formulated as follows:

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{T}\boldsymbol{P}^\top \\ \boldsymbol{Y} &= \boldsymbol{U}\boldsymbol{Q}^\top \end{aligned} \tag{2.3}$$

where $\boldsymbol{P}$ and $\boldsymbol{Q}$ are orthogonal matrices of loadings and $\boldsymbol{T}$ and $\boldsymbol{U}$ are the projections of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. The model aims at maximising the covariance between the scores $\boldsymbol{T}$ and $\boldsymbol{U}$.

Neural networks Neural Networks (McCulloch and Pitts, 1943) are non-linear supervised ML models. Neural Networks are constructed as chains of functions composed together. For example, we might have three functions $f^{(1)}, f^{(2)}$, and $f^{(3)}$ connected in a chain to form $f(x) = f^{(3)}(f^{(2)}(f^{(1)}(x)))$. The function f is a three-layers neural network, and each $f^{(i)}$ represents a network layer. Generally,

the functions $f^{(i)}$ take the form $f^{(i)}(\boldsymbol{x}) = g(\boldsymbol{W}^i\boldsymbol{x} + \boldsymbol{b}^i)$, where $\boldsymbol{W}^i$ and $\boldsymbol{b}^i$ are the layer parameters and g a non-linear function such as the Hyperbolic tangent.

Multiple Neural Network classes have been proposed, and we are especially interested in Convolutional Neural Networks (CNNs). CNNs are specialised neural networks for processing data with a known, grid-like topology of the input data, such as images. A CNN learns the parameters of a set of filters in each layer, which are used to convolve the input and pass its result to the next layer. After passing through convolutional layers, the input data becomes abstracted to feature maps. Additionally, neural networks can be built as Encoder-Decoders, a Neural Network structure that learns to reconstruct the input data. Internally, it has a hidden layer representing the latent space from which the input data were generated. These structures are commonly used for image segmentation and dimensionality reduction.

Neural Networks constitute a flexible family of models that can be designed and adapted to a wide range of tasks, including regression, classification, segmentation and survival prediction. However, complex and deep Neural Networks are associated with a large number of parameters that must be estimated. Therefore, a populous dataset is required in order to meaningfully learn using Neural Networks.

2.3.3.2 Survival models

Survival analysis is the estimation of the expected duration of time until one event occurs. For example, survival analysis in the oncological context includes the estimation of the duration from the initiation of treatment to the occurrence of disease progression or death (progression-free survival). Generally, instead of trying to estimate the event time directly, most survival models estimate a hazard function. A hazard function relates the passage of time and other covariates with the probability of an event occurring. Unlike typical regression problems, in which all training output variables are known, survival problems differ in that only some of the events in the training dataset are observed while the others are *censored*. In particular, several patients might not have a death date but only the last consultation date.

Cox-model The Cox proportional hazard (Cox, 1972) models the hazard function as the multiplication of a time-dependent baseline function ($\lambda_0(t)$) and an exponential risk function which depends only on the individual's covariates. The hazard function (associated with the risk r), at a time τ and individual's covariates $\boldsymbol{x}$, is given as:

$$\begin{aligned} h(\boldsymbol{x}, \tau) &= \lambda_0(t) exp(\boldsymbol{x}^T\boldsymbol{\beta}) \\ r(\boldsymbol{x}) &= exp(\boldsymbol{x}^T\boldsymbol{\beta}) \end{aligned} \tag{2.4}$$

The parameters $\boldsymbol{\beta}$ can be estimated without considering the time factor. Given n subjects with covariates $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n$ and survival times $(t_1, c_1), \cdots, (t_n, c_n)$ with t_i being event times and c_i being censoring times, we want to maximise the probability that individual j has an event at $\tau_j = min(t_j, c_j)$ compared to the other individuals still at risk at τ_j. We note $\mathcal{R}_j = i \in [1, n] | \tau_i \geq \tau_j$, and we formulate our target likelihood for individual j:

$$L_j = \frac{exp(\boldsymbol{x}_j\boldsymbol{\beta})}{\sum_{i\in\mathcal{R}_j} exp(\boldsymbol{x}_i\boldsymbol{\beta})} \tag{2.5}$$

When multiple individuals have the same event time, Breslow (1975) proposed regrouping them as having the same risk at τ_j, and the reasoning shifts from the individual to the event time. Given D_j individuals having the same survival time τ_j, the likelihood function becomes for τ_j:

$$L_j = \prod_{j\in D_j} \frac{exp(\boldsymbol{x}_j\boldsymbol{\beta})}{\sum_{i\in\mathcal{R}_j} exp(\boldsymbol{x}_i\boldsymbol{\beta})} \tag{2.6}$$

The parameters $\boldsymbol{\beta}$ are obtained by maximising the log-likelihood of observing all events:

$$\mathcal{L} = log\left(\prod L_j\right) \tag{2.7}$$

When covariates exceed the number of individuals, the log-likelihood is penalised using ElasticNet (Simon et al., 2011a). Given α and r two positive numbers, with $r \in [0, 1]$, the loss function is defined as:

$$\mathcal{L}_{penalised} = -\mathcal{L} + \alpha\left(r||\boldsymbol{\beta}||_1 + \frac{1-r}{2}||\boldsymbol{\beta}||_2^2\right) \tag{2.8}$$

Random-forest Random forest (Ho, 1995) is a machine-learning model based on the ensembling of multiple decision trees. Each tree is built by recursively splitting the dataset according to a split criterion. Individual trees are ensured to be uncorrelated by building each ones on a different bootstrap sample of the original data and by evaluating the split criterion only for a randomly selected subset of the variables.

In the case of survival prediction, multiple splitting criteria have been proposed. We chose to use the Log-rank test. This is a non-parametric hypothesis test to compare the survival distributions of two samples. Log-rang test estimates, under the null hypothesis Ho, the probability that two samples of individuals have such an extreme difference in hazard functions. The chosen variables must minimise the Log-rank probability.

2.3.3.3 Unsupervised models

Having observed data $\boldsymbol{X}$, unsupervised models focus on learning patterns in the data without a target at hand. This section focuses on some of the most commonly used dimension reduction models.

Non-negative Matrix Factorisation Non-negative Matrix Factorisation (NMF) (Lee and Seung, 1999) aims at finding two non-negative (all coefficients are non-negative) matrices $\boldsymbol{W}$ and $\boldsymbol{H}$ of sizes n, k and k, p respectively, such that $\boldsymbol{X} \approx \boldsymbol{W}\boldsymbol{H}$. $\boldsymbol{W}$ ($\boldsymbol{H}$) contains a latent representation for the n samples (p covariates) in a k-dimension space. This is done by minimising the loss function $||\boldsymbol{X} - \boldsymbol{W}\boldsymbol{H}||_F$ under the constraints of non-negativity for $\boldsymbol{W}$ and $\boldsymbol{H}$.

Principal Component Analysis Principal Component Analysis (PCA) (Pearson, 1901) is a dimensional reduction method. It linearly transforms the data $\boldsymbol{X}$ into an orthogonal coordinate system. Data is then projected into a smaller dimension where most variance is explained. This is done by diagonalising the estimated covariance matrix ($\boldsymbol{X}^\top \boldsymbol{X}$), which is a symmetric semi-positively defined matrix.

Canonical Correlation Analysis Canonical Correlation Analysis (CCA) (HOTELLING, 1936) is a method for inferring information from cross-correlation matrices. Having two observation matrices $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, the model projects the blocks into a latent space where their correlation is maximised. Formally, the model finds the optimal projectors $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}$ that maximise the function: $corr(\boldsymbol{X}^{(1)}\boldsymbol{w}^{(1)}, \boldsymbol{X}^{(2)}\boldsymbol{w}^{(2)})$.

2.3.4 Metrics

The performance of statistical models can be assessed through multiple methods. Here we present the main metrics used in this thesis.

2.3.4.1 Binary classification metrics

Having a binary prediction $\hat{\boldsymbol{Y}}$, we aim to assess its agreement with the binary ground truth $\boldsymbol{Y}$. The two can be viewed as sets of positive labels. Five metrics are commonly used.

The recall measures the fraction of positive labels retrieved by the prediction and is defined as :

$$Recall(\hat{\boldsymbol{Y}}) = \frac{|\hat{\boldsymbol{Y}} \cap \boldsymbol{Y}|}{|\boldsymbol{Y}|} \tag{2.9}$$

The precision measures the fraction of correctly predicted positive labels among all predicted labels.

$$Precision(\hat{\boldsymbol{Y}}) = \frac{|\hat{\boldsymbol{Y}} \cap \boldsymbol{Y}|}{|\hat{\boldsymbol{Y}}|} \tag{2.10}$$

The dice or f-score is an aggregation of the precision and recall metrics. It is defined as

$$Dice(\hat{\boldsymbol{Y}}) = \frac{2 \times |\hat{\boldsymbol{Y}} \cap \boldsymbol{Y}|}{|\hat{\boldsymbol{Y}}| + |\boldsymbol{Y}|} \tag{2.11}$$

For binary classification, the model returns a probability map for each class. Thus, these probability maps are first thresholded (typically at 0.5) to obtain the prediction. Recall, Precision and Dice all depend on the threshold. The Area Under Curve (AUC) is a metric independent of the threshold by integrating the precision-recall curve. A 0.5 value indicates random results.

2.3.4.2 Survival metrics

The concordance index (C-index) measures the survival model ability to correctly rank the survival times based on the individual risk scores. The C-index generalises the area under the curve, considering the censored data. A higher C-index is better, with 0.5 indicating random results and 1 reflecting a perfect ranking of the individuals. Let δ_i be a binary variable indicating if the survival time of

individual i is non-censored ($\delta_i = 1$ if survival time is not censored, 0 otherwise), η_i his estimated risk score, and T_i his observed event time, the C-index is given by:

$$\begin{aligned} \text{C-index} = & \frac{\sum_i \sum_j 1_{T_j<T_i} \cdot 1_{\eta_j>\eta_i} \cdot \delta_j}{\sum_i \sum_j 1_{T_j<T_i} \cdot \delta_j} \\ & \text{with:} \\ & \quad \eta_i, \text{the risk score of a patient } i \\ & \quad \mathbf{1}_{T_j<T_i} = 1 \quad \text{if} \quad T_j < T_i \quad \text{else} \quad 0 \\ & \quad \mathbf{1}_{\eta_j>\eta_i} = 1 \quad \text{if} \quad \eta_j > \eta_i \quad \text{else} \quad 0 \end{aligned} \tag{2.12}$$

* * *
* *
*

Chapter 3

Related Works

In this thesis, we are interested in integrating radiomic with multi-omic data. This integration is intended to reveal associations between non-invasive data and molecular data, which can improve our knowledge of the pathologies at hand. As a first step, it is necessary to position our project in the context of its related works.

We start by reviewing the various studies involving radiomic analysis. We are particularly interested in the tumours we want to study, namely Lower-Grade Gliomas, Glioblastomas and DIPG. Since radiomic feature analysis relies on a precise Region of Interest identification, we will review brain tumour segmentation methods proposed in the literature. We are interested in the best-performing methods in international challenges. Later, we review different methods proposed to integrate multi-source data, distinguishing between supervised and unsupervised approaches. Finally, we review several strategies to help understand and interpret the built models.

Chapter Outline

Contents

3.1 Radiomics 23
3.1.1 Classification of tumour types or subtypes using radiomics 23
3.1.2 Survival Prediction using radiomics 23
3.1.3 Reproducibility of radiomics 24
3.2 Brain tumour segmentation 26
3.2.1 Using custom features 26
3.2.2 Using Deep-learning 27
3.2.3 Tackling issues with brain tumour segmentation 28
3.3 Multi-block approaches 29
3.3.1 Unsupervised models 29
3.3.2 Supervised models 31
3.4 Variable selection 32
3.4.1 Sparsity-based penalties 32
3.4.2 Graph based penalties 33

3.1 Radiomics

Several studies have shown the effectiveness of radiomic features in tumour analysis. These studies include but are not limited to, the classification of tumour types or subtypes and disease prognosis. Often, the radiomic features are extracted from a region of interest in the image and then used in a machine learning or statistical model to make a prediction.

3.1.1 Classification of tumour types or subtypes using radiomics

Malik et al. (2021) studied the ability of radiomic features to distinguish between LGG and Glioblastomas. On 74 patients (32 LGG and 42 Glioblastoma), the authors report an AUC score between 0.72 and 0.96, depending on the feature selection method and the classification algorithm. The most recurrent selected features are texture ones, especially from the CE T1w. However, their results show a lack of stability in variable selection: most features are chosen only by one method and on less than 70% of the runs.

Kong et al. (2019) studied the O^6-methylguanine-DNA methyltransferase (MGMT) methylation status in patients with primary gliomas as classification target. On 107 patients (59 methylated and 48 unmethylated), the authors used radiomic features extracted from positron emission tomography (PET) images to classify the methylation status. They report an AUC of 0.86 on the test set. Comparatively, the clinical data achieved an AUC of 0.69, and the combination of radiomic and clinical data gave an AUC of 0.85. Most variables of interest were first-order radiomics.

3.1.2 Survival Prediction using radiomics

3.1.2.1 Radiomics on the glioblastoma

Kickingereder et al. (2016) studied the performance of radiomics extracted from MR images when used to predict survival in 119 glioblastoma patients. The authors report a c-index of 0.65 when predicting overall survival with radiomics alone and 0.70 when clinical data were added to the radiomics. Comparatively, clinical data reached a c-index of 0.64. In this study, all variables of interest were texture features extracted from the FLAIR modality.

Kickingereder et al. (2017) analysed the added value of radiomics (extracted from MR images) to clinical and key molecular data for disease stratification. The study included 181 glioblastoma patients and reports that radiomics alone perform similarly to clinical data and better than molecular data. Combining the information from multiple sources reduced the prediction error compared to each single-source model. The highest accuracy model was built using data from all sources (i.e., clinical + molecular characteristics + radiomic features). Their approach identified 8 radiomic features that were the most important for predicting the outcome. These features included texture features (6/8 features derived from FLAIR and T2w) and volumetric features (2/8 features derived from CE T1w).

3.1.2.2 Radiomics on the LGGs

Radiomics have also been used on LGGs. For instance, Liu et al. (2018) studied the progression free survival of LGGs using radiomic features extracted from T2w MR images and compared the results

obtained with clinical and histological data. The study found that first-order features are the most relevant for the problem at hand, and these radiomic features outperformed the other data sources.

Wagner et al. (2021) studied the ability of radiomic features to discriminate between two mechanisms of implication (gene fusion or gene mutation) of the B-Raf proto-oncogene, serine/threonine kinase (***BRAF***) in pediatric LGGs. The study extracted the radiomic features from the FLAIR sequence on 115 patients and reports that the most important features were texture related, extracted from wavelet-transformed images. These features were successfully used to stratify the patients with a ROC-AUC of about 0.75.

3.1.2.3 Radiomics on the DIPG

Some studies, although few in number, have examined the DIPG using radiomics. For example, Tam et al. (2021) studied the survival prediction using radiomics from the T1w and T2w on 153 patients. The study isolated five features as the most relevant. These features included three features from T1w and two features from T2w. All features were intensity and texture based on the wavelet-filtered images. The authors report that radiomics outperformed the clinical model that used sex and age at diagnosis as variables (a C-index of 0.55 for radiomics and 0.51 for clinical). When clinical features were combined with radiomics, the model performance increased but not significantly over radiomics alone. The performance of the radiomics features considered separately from sequences T1w and T2w was lower compared to their combination. Additionally, their performances were not significantly better than the random predictions, with an average C-index of 0.51 and 0.55 when using the T1w only and the T2w only, respectively.

The reviewed articles show that radiomic are promising predictors for various tasks on different tumours. However, they also suggest stability issues, probably due to the important co-linearity among radiomic variables. The studies do not agree on the most critical modalities or predictive features. No two studies have reported the same variables of interest. We note, however, that these studies are done in comparatively small studies, and further studies must be done to confirm the results.

3.1.3 Reproducibility of radiomics

Despite the successes of radiomic analyses in oncological studies, several factors influencing the stability and reliability of the feature extraction procedure have been identified in the literature (van Timmeren et al., 2020). For MR scans, these factors include image acquisition, resolution, reconstruction, preprocessing, and region of interest (ROI) delineation.

Um et al. (2019) studied the effect of the MRI field strength and manufacturer on texture radiomics. The authors studied 50 patients from The Cancer Genome Atlas (TCGA)-GBM. These patients met the following inclusion criteria: newly diagnosed GBM, pre-surgery status, and availability of multimodal MRI, including the T1w, FLAIR and CE T1w. The authors compared 32 scans acquired with a 1.5T with 13 with 3T scanners (they discarded other magnetic field strengths). Using the Wilcoxon test, the authors found that around 14% of GLSZM features and 7% of GLCM features were manufacturer dependent; and around 28% of GLSZM and 20% of GLCM features were field-strength dependent. The authors conclude that there is a bias introduced during the image acquisition.

Molina et al. (2017) investigated the effects of the MR image resolution, slice thickness and dynamic grey-level range on 16 texture radiomic features extracted from Glioblastoma scans. The study was done on 20 patients who had T1w pre-treatment scans available. For each patient, the authors compared two matrices of the same slice: 432x432 (raw matrix) and 256x256 (standard MRI matrix size). They compared the features using the coefficient of variation, a standardised measure of dispersion defined as the ratio between the standard deviation and the mean of a series of data. The authors found that the investigated radiomic features are not robust to resolution and slice thickness. Only the entropy extracted from the co-occurrence matrix resisted the robustness test regarding the dynamic range.

The dependence of texture indices with the image pre-processing (multiple sequence image alignment, intensity normalisation, level binning and ROI generation) was studied on about 30 patients with DIPG by (Goya-Outi et al., 2018). A task of classification of control tissue (white matter) versus tumour tissue was considered. Although the study reports AUCs above 0.80, it also shows that image preprocessing impacts the results, particularly binning, which is highly dependent on inter-site image harmonisation.

Brynolfsson et al. (2017) studied the effect of five imaging and pre-processing parameters on the GLCM texture features. These parameters are: the noise level, the resolution, the ADC map construction method, the quantisation method, and the number of grey levels in the quantized images. The authors found that most features are affected by the investigated parameters using two datasets, glioblastoma and prostate cancer. Only the b-values used for constructing the ADC maps in the glioma data set had no significant effect on any feature.

Moradmand et al. (2020) studied the effects of bias correction and noise reduction on the radiomic features. Overall, the total number of high-robustness features extracted from the necrosis region (30.6%) was higher than the number of features extracted from oedema regions (20.2%), enhancement regions (19.2%), and active tumour regions (17.3%). Additionally, the average number of highly reproducible features for baseline comparison with bias field correction (23.2%) was higher than for baseline comparison with noise correction (21.4%). Interestingly, performing the bias field correction before noise correction increased the stability (22.5%) compared with the noise correction before bias correction (20.4%).

Poirot et al. (2022) studied the effect of segmentation on radiomic features. They used four classical segmentation software to compare radiomic features extracted from several brain compartments. The results show that despite a high segmentation agreement between the different methods (all above 0.75%), the radiomic features present significant differences. This is especially true for texture features extracted from T1w scans.

Peerlings et al. (2019) studied the reproducibility of radiomic features. Scans were acquired twice within seven days under similar conditions, and their radiomic features were compared using the Concordance Correlation Coefficient (CCC). The results show that only around 25% of the features were reproducible. Nevertheless, radiomics features achieved comparable reproducibility after wavelet-filtering, which can alleviate boundary inhomogeneity. The authors also found that features extracted from 3T images were more stable (32%) compared to 1.5T images (25%). No significant difference in stability was observed when comparing the scanner manufacturers; however, the stable features have only around a 58% overlap. Finally, 122 features were identified as stable across multiple cancer types

(ovarian, lung and liver cancer).

The usage of radiomic features extracted from MR scans is still emerging compared to CT and PET scans. Thus, studies of the different stability issues, solutions, and protocols to alleviate these instabilities are still developing. Cattell et al. (2019) and van Timmeren et al. (2020) present multiple works on the topic. The presented works suggest that radiomic features can be good predictors and can help in learning about various diseases. However, these features are not stable or reproducible to be used as biomarkers for tumour types.

3.2 Brain tumour segmentation

Image segmentation is a key step for radiomic analysis. Identifying the Region of Interest (ROI) from the MR scan is important to extract meaningful descriptors of the tumour at hand. Given an input volume from one or multiple modalities, automatic brain tumour segmentation refers to the differentiation between the tumour (and its sub-compartments) from healthy brain tissue. This is done by statistical models that map each voxel into a set of sub-compartment labels.

In this section, we discuss the various automatic brain tumour segmentation methods. We focus on the best-performing models reported in the Brain Tumour Segmentation (BraTS) challenge (Menze et al., 2015). The challenge is focused on the segmentation of Glioblastoma and LGGs. The challenge has occurred since 2012 and reflects advances in the automatic brain tumour segmentation problem.

3.2.1 Using custom features

Random Forests have been one of the early successful models for brain tumour segmentation. For instance, Criminisi et al. (2012) successfully used a Random Forest model to achieve a top rank in the BraTS challenge 2012. Their model segments the image into background, tumour and oedema by independently assigning a label to each voxel. First, given an intensity, they estimate its probability of belonging to each class. Then, these probability maps were combined with the original MR scans and used as input to their random forest. Their method achieved a Dice score of 0.7 on the whole tumour and 0.2 on the tumour core.

Reza and Iftekharuddin (2013) also used a Random Forest as an automatic brain tumour segmentation model. From the available MR scans, they computed the difference in voxel intensities between the different modalities, and they extracted the fractal Piece-wise Triangular Prism Surface Area (PTPSA) (Islam et al., 2013) and textons (Leung and Malik, 2001). Then they combined these features with the MR scan intensities and trained their Random Forest. This method was among the highest-ranked models in BraTS 2013, with a reported Dice of 0.83 on the whole tumour and 0.72 on the tumour core.

Dvorak and Menze (2015) proposed to redefine the segmentation output. The authors separated the segmentation problem into three binary segmentation sub-tasks: the whole tumour, the tumour core and the enhancing tumour segmentation. For each sub-task, they divided the label image into binary patches. This is done to extract a set of possible local appearances of each label, which they reduced into a few templates using k-means. Finally, the authors used Convolutional Neural Networks (CNN) to map the original image patch into its closest template, and all overlapping predictions of a

neighbourhood are averaged. These predictions are thresholded at 0.5 to obtain the final prediction label. Using this method, the authors obtained a Dice score of 0.83, 0.75 and 0.77 on the whole tumour, the tumour core and the enhancing tumour, respectively.

3.2.2 Using Deep-learning

Since 2015, almost all best-performing models in the BraTS challenge have used Convolutional Neural Networks (CNN). Here, we discuss the multi-path networks and encoder-decoder framework. These two classes of CNNs have been prominent in recent research about automatic brain tumour segmentation.

3.2.2.1 Multi-path networks

A neural network path is defined as a chain of data processing. Most Multi-paths neural networks divide the input image into small patches and aim at finding the label of the central pixel or voxel at the centre of each patch. Using multiple paths allows the analysis of the patch at different scales. Thus utilising local details and global information to make a prediction about the target voxel.

Havaei et al. (2017) proposed a two-path 2D CNN-based model to classify the central pixel of each patch. Their approach only uses the axial view as it has the best resolution. The model combines the results of two simultaneously trained CNNs, each using a different patch size. The chosen sizes were 33x33 and 65x65. The output of the first CNN is concatenated with the second CNN input, thus treated as additional channels of the input patch. The local path uses 7x7 convolution kernels, while the global path uses 13x13 convolution kernels. Finally, the concatenation of the feature map is fed to the last output layer of the model to make a prediction. The authors achieved second rank in the 2015 BraTS challenge with an average reported dice of 0.88 on the whole and 0.79 on the core tumour.

Kamnitsas et al. (2016) proposed DeepMedic an 11-layers deep, multi-scale 3D CNN. The model is built using two parallel paths, each processing the input at a different scale. The low-resolution and high-resolution paths take as input 3D patches of sizes 19x19x19 and 25x25x25, respectively. The network segments a small patch of 9x9x9 at the centre of the input data. The authors also propose to extend the original model with residual connections (adding the results of earlier layers as input to some later layers) to help train the network by facilitating signal preservation. The model was one of the top-performing networks during BraTS 2016, with a reported dice of 0.89 on the whole tumour and 0.76 on the tumour core. The addition of the residual connections had a small improvement in the results.

Similarly to Havaei's model, Ben naceur et al. (2020) proposed CNN architectures designed to predict a central pixel class using the 2D patch from the image. They propose to use overlapping patches to help the model utilise local and global information. Thus, they used five predictions per patch, where the prediction of the adjacent patches influences the results obtained on the current patch. The authors report a dice score of 0.90 on the whole tumour and 0.83 on the core tumour when tested on the BraTS 2018 data.

3.2.2.2 Encoder-decoder networks

Using multi-path networks to propose a classification of the central pixel (voxel) of a patch is computationally challenging. There is a lot of computation redundancy, increasing training and inference times. Furthermore, the accuracy of the prediction can be heavily impacted by the patch and its source quality and size. To tackle these issues, models have been proposed to perform end-to-end segmentations. These methods encode the whole image into a latent representation, which is then decoded into the segmentation mask of the same size as the input.

The most popular encoder-decoder architecture is the UNet (Ronneberger et al., 2015). UNet is a symmetric Encoder-Decoder CNN. The model has skip connections between the encoding part and the decoding part. These skip connections directly pass features from a latent encoding layer to its symmetric decoding layer, which helps the model recover details. Casamitjana et al. (2016) used the UNet architecture in the BraTS challenge 2015 and reported a dice score of 0.89 on the whole tumour and 0.76 on the core tumour.

Isensee et al. (2018) proposed a modified UNet to perform brain tumour segmentation. To do so, the authors used feature maps from the decoding path to generate secondary segmentation maps. These are combined with the final segmentation, and the losses associated with these segmentation maps are added to the final loss function. The authors also used a dice loss function to tackle class imbalance and performed data augmentation. The model was one of the top-ranked during the BraTS 2017 challenge, with a dice score of 0.90 and 0.80 on the whole and core tumour, respectively.

Milletari et al. (2016) proposed V-net, a modified UNet. V-net defines the data flow as a chain of multiple stages operating on different resolutions, each stage comprising one to three convolution layers. V-net modifies UNet by adding the input of each stage in the encoding path to the output of the last convolutional layer of that stage in order to enable learning a residual function. Hua et al. (2020) successfully used V-net for brain tumour segmentation. The authors report a dice score of 0.88 on the whole tumour and 0.8 on the tumour core when tested during the BraTS 2018 challenge.

3.2.3 Tackling issues with brain tumour segmentation

Brain tumour segmentation corresponds to a classification task with very unbalanced classes. This means that not all classes are represented equally: the tumour generally occupies 15% of the brain, of which the œdema represents generally 70% of its appearance. Additionally, using promising neural networks requires a large set of training data, which is often hard to satisfy. Multiple approaches have been proposed to tackle these issues. Here we focus on three main solutions: (a) training the models on multiple tasks (multi-task approaches), (b) customising the loss function, and (c) using priors.

3.2.3.1 Multi-task approaches

Multiple works introduced auxiliary tasks that are chained to enhance the model performance. The underlying idea is that tasks share information. Thus, the easier task helps solve the harder task. In the case of neural networks, this is done by making all the tasks share the earlier extracted features.

Zhou et al. (2019) used an architecture similar to UNet. The authors proposed to train their model on three tasks simultaneously: the binary segmentation of the whole tumour, the detailed

segmentation of the tumour, and the binary segmentation of the enhancing tumour. When tested on the BraTS 2018 challenge, the authors report an average Dice of 0.88 on the whole tumour and 0.80 in the tumour core.

Myronenko (2018) proposed NV-net, an encoder-decoder architecture with an additional branch to the decoder part to reconstruct the original image. When tested on the BraTS 2018 challenge, the authors report an average Dice of 0.88 on the whole tumour and 0.82 in the tumour core.

3.2.3.2 Using priors

Some works added prior knowledge to help their model learn helpful features for the brain tumour segmentation task. For instance, since the tumours appear differently on the FLAIR and CE T1w, Longwei and Huiguang (2018) proposed using two models, each using a different modality. The first used the FLAIR to segment the whole tumour, and the second used only the CE T1w to segment the tumour core. When tested on the Brats 2018, the authors reported a Dice score of 0.85 and 0.72 on the whole tumour and core, respectively.

Rosana et al. (2020) proposed Bounding Box UNet (BB-UNet), a neural network model that expends the UNet model to consider prior bounding boxes. The proposed model incorporates priors through novel convolutional layers introduced at the level of skip connections to guide the model on where to look for. It must be noted that the bounding boxes have to be defined beforehand.

3.3 Multi-block approaches

Let $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \cdots, \boldsymbol{X}^{(J)}$ be J blocks, representing measures of different sources on the same n samples. Multi-block analysis frameworks are well established by numerous works that proposed to adapt classical statistical models for the multi-block setting. These methods often involve adapting the loss functions or the learning strategy. In this section, we discuss different methods that we think of interest.

3.3.1 Unsupervised models

Different approaches have been proposed to analyse multi-block structured data in an unsupervised manner. These models generally aim at a joint dimensionality reduction, thus getting a more comprehensive overview of the studied context. Here, we present the most relevant unsupervised methods from our perspective.

3.3.1.1 Models built on the NMF

The Non-negative Matrix Factorisation (NMF) has been extended to the multi-block framework. Joint NMF (jNMF) (Zhang et al., 2012) defines multiple NMF problems with a shared factor matrix across all blocks. This method is equivalent to the classic NMF with the concatenation of the blocks into a single matrix.

Integrative NMF (iNMF) (Yang and Michailidis, 2015) is another method based on NMF. Similarly to the jNMF, iNMF defines a set of NMF problems. However, instead of only using a single shared

factor matrix, the latent factor matrix is decomposed into a shared and a block-specific factor matrix. jNMF and iNMF can be seen as a simultaneous clustering of observations and variables.

The Multi-Omics Factor Analysis (Argelaguet et al., 2018) (MOFA) is another unsupervised method for decomposing the blocks into the products of a shared factor matrix and block-related weight matrices. The MOFA defines a probabilistic Bayesian problem, where prior probabilistic distributions are imposed on all the estimated variables. Instead of the least-square approximation used by the NMF, MOFA uses a variational inference to estimate the parameters. To make the model more interpretable, the authors propose adding sparsity constraints by using Automatic Relevance Determination (ARD) and spike-and-slab priors on the factor and weight matrices. This is done by the multiplication of the priors on the factor and weight matrices by a set of random variables having a Bernoulli distribution; each prior will select the variables of interest (or make them zero).

3.3.1.2 Models built on the PCA

The Multiple Factor Analysis (MFA) (Escofier and Pagès, 1994) is an extension of the classical factor analysis method to the multi-block framework. The MFA can be described as a Principal Component Analysis (PCA) of the concatenated blocks. Each block is weighted using the largest eigenvalue from its associated correlation matrix. The weighting ensures a single block variance or dimension does not dominate the global PCA first component.

The Joint and Individual Variation Explained (JIVE) (Lock et al., 2013) decomposes each block into the sum of two low-rank matrices. The first matrix represents the joint structure matrix, which contains common information between all the blocks. The second matrix represents the individual information of each block. The joint and individual matrices are required to be column-orthogonal. JIVE can be seen as two successive PCA decompositions: the first decomposition extracts information from the global structure (using the concatenation of all the blocks), and the second PCA is done on the residualized individual blocks.

3.3.1.3 Models built on the CCA

The Canonical Correlation Analysis (CCA) is a statistical tool to find shared information from two blocks by maximising their correlation in a latent space. Multiple methods have been proposed to extend the CCA to take into account more than two blocks. These methods are referred to as Generalized CCA (GCCA). Formulations of the GCCA include SUMCOR, MAXVAR, among others (Horst, 1961; KETTENRING, 1971). The GCCA has been further expanded to include regularity or sparsity in the model Regularized GCCA (RGCCA) (Tenenhaus and Tenenhaus, 2011) and Sparse GCCA (SGCCA) (Tenenhaus et al., 2014). The RGCCA and SGCCA will be detailed in later chapters.

Another method is the Multiple CO-inertia Analysis (MCOA) (Bady et al., 2004), which extends the Co-inertia Analysis to a set of blocks. This is done by the optimisation of the covariance between several individual ordinations and a shared reference ordination.

Other methods have been proposed to analyse multi-blocks requiring special block configurations. For example, the Tensorial independent component analysis (tICA) (Teschendorff et al., 2018) is an extension of the PCA for tensorial data. However, this algorithm is designed for datasets with blocks that share the same samples and variables. The Data Fusion algorithm (Zitnik and Zupan, 2015)

requires relationship matrices between pairs of blocks. A comparison between several discussed methods has been presented by Cantini et al. (2021). The study concludes that the different benchmarked algorithms have different results depending on the studied issue. For instance, for clustering tasks, iNMF performed the best as designed for this task, while JIVE and the RGCCA had the best performance among the set of methods not intrinsically designed for clustering. Furthermore, MCIA, JIVE, MOFA, and RGCCA were the most efficient for detecting latent features associated with survival.

3.3.2 Supervised models

Multiple approaches have been proposed to extend classic regression models to account for separate blocks, each having distinct variables. Two main strategies can be identified: combining multiple models, each learned from a single block, and learning a unique joint model from the different blocks.

3.3.2.1 Combination of multiple models strategies

Priority-Lasso (Klau et al., 2018) is a multi-block regression model based on multiple regression models, each trained on a separate block. Priority-Lasso requires user-defined order of the blocks, which defines their priority to the model. The algorithm sequentially fits regression models on each block, using the obtained results from the previous block as an offset. The fitted models can be over-optimistic on the training set, making the offset capture variability not contained by the model. To tackle this issue, the algorithm uses a cross-validation approach. On an acute myeloid leukaemia dataset, it showed similar results in terms of prediction accuracy compared to a Lasso model. The authors also showed a dependency of the model on the user-defined priority order.

Sequential and Orthogonalized-PLS (SO-PLS) (Jørgensen et al., 2007) is a regression model based on the Partial Least Square (PLS) regression. Given two blocks, the algorithm starts by fitting the output to the first block using a PLS regression. Then, it uses the extracted PLS scores to orthogonalise the second block before using it in a second PLS regression. Finally, the model combines the two computed PLS scores and uses a classical least square regression to predict the output. The authors claim that the impact of the blocks order is minimal regarding the prediction. The method can be easily expanded to multiple blocks.

Parallel Orthogonalised-PLS (PO-PLS) (Måge et al., 2008) is another multi-block regression model based on the PLS. Instead of sequentially regressing the different blocks, the authors propose to extract shared information first. This is done by computing PCA scores from the blocks, then using the Canonical Correlation Analysis (CCA) (or the Generalised CCA when more than three blocks are involved) on the PCA scores to obtain the canonical coefficients. Only the components having a canonical correlation exceeding 0.95 are kept. The original PCA scores are then orthogonalised against the canonical coefficients. Finally, a PLS regression model is built using the common and orthogonalised components.

3.3.2.2 Structured penalties strategies

Structured penalties have been proposed to integrate multiple blocks into machine learning models. For example, Integrative LASSO with Penalty Factors (IPF-LASSO) (Boulesteix et al., 2017) proposed

using a weighted sum of the ℓ_1 norms of the coefficient vectors of each modality as a penalty. The proposed penalty adapts the sparsity for each block, which allows adjusting the proportion of variables of interest selected from each modality. However, this comes at the cost of the number of parameters needed to tune, which is proportional to the number of blocks at hand.

Random forests have been expanded to take into account the multi-block structure of the data in an approach called BlockForest (Hornung and Wright, 2019). The classic random forest algorithm takes a small sample of the variables on which it computes a criterion and makes a splitting decision. The imbalance of the number of variables from each block causes an over-representation of variables from larger blocks (in terms of the number of variables), regardless of their link to the outcome. To tackle this issue, the authors propose five sampling strategies. In VarProb, each variable has a block-specific variable sampling probability. With SplitWeights, the split criterion is weighted according to the block. Using BlockVarSel, the algorithm samples a fixed number of variables from each block. In RandomBlock, each split is done on a single, randomly selected block. Finally, with BlockForest, a random set of blocks is chosen before BlockVarSel is applied.

Other methods based on classic machine learning strategies, such as gradient boosting, have been proposed to take into account the block structure of data. Neural networks have also been used for supervised multi-block learning, and a study of these methods can be found at Leng et al. (2022). A benchmark of some of the discussed methods has already been done by Herrmann et al. (2020) in the context of survival prediction with multi-omics. When tested on the survival prediction task, the study reports that BlockForest was the best-performing model among all tested models. However, the study also reports a high variability of the model performances when tested on different tumour types. For instance, Priority-Lasso was the best-performing model when tested on liver, lung and ovarian cancers, but its performance also dropped below the random results on other cancers.

3.4 Variable selection

The usefulness of statistical models is not limited to their interference performance but also their ability to exhibit variables of interest. Variable selection is an essential regularisation tool to fit models when the number of studied variables far exceeds the number of samples available, which is the most frequent case when analysing biological data. As a side-effect, this allows better interpretability of a built model and the identification of the most critical variables, which opens further investigation into them. In this section, we will discuss the different proposed methods to perform variable selection.

3.4.1 Sparsity-based penalties

One of the most common approaches for variable selection is introducing a sparsity constraint, which refers to setting the coefficients of non-important variables to zero. Most of these methods are adaptations of the Least Absolute Shrinkage and Selection Operator (LASSO) (Tibshirani, 1996). Almost all of these methods focus on respecting or exploiting dependencies among the studied variables.

The Fused Lasso is another adaptation of the Lasso penalty proposed by Tibshirani et al. (2005). The penalty was designed for ordered data such as time series. The penalty aims at selecting a continuous sequence of variables. The penalty takes the form $\lambda_1||\boldsymbol{\beta}||_1 + \lambda_2 \sum |\beta_i - \beta_{i+1}|$. The authors

proposed to estimate the order from the data when there is no *a priori* order known. For example, they organised similar variables near one another using hierarchical clustering and applied it to gene expression data.

Yuan and Lin (2006) introduced the Group Lasso penalty, which allows a model to jointly select (or discard) all variables within the same group. The groups are predefined and reflect *a priori* knowledge. The penalty takes the form $\lambda_1 \sum_{j=1}^{J} ||\boldsymbol{\beta}_j||_{\boldsymbol{K}_j}$ with $\boldsymbol{K}_j$ a symmetric positive definite matrix and the estimates $\boldsymbol{\beta}$ have been separated into J disjoint groups.

The Octagonal Shrinkage and Clustering Algorithm for Regression (OSCAR) has been proposed by Bondell and Reich (2007) and aims at simultaneously selecting variables while grouping them into predictive clusters. OSCAR encourages the grouping of highly correlated covariates. This is done by combining a Lasso penalty and a pair-wise ℓ_∞. The penalty takes the form $\lambda_1||\boldsymbol{\beta}||_1 + \lambda_2 \sum_{j<k} max(|\beta_j|, |\beta_k|)$. While the two previously discussed penalties rely on *a priori* knowledge, OSCAR infers the groups from the data.

Clustered Lasso, by She (2010), is another adaptation of the Lasso penalty. Similar to OSCAR, the penalty aims to group similar variables of interest into clusters without *a priori* knowledge. The penalty is built as a combination of the Lasso and a pairwise absolute difference between all the coefficients. The penalty takes the form $\lambda_1||\boldsymbol{\beta}||_1 + \lambda_2 \sum_{j<k}(|\beta_j - \beta_j|)$. While Clustered Lasso and OSCAR seem similar, they differ in their treatment of anti-correlated variables. The authors found that while the predictive performance of Clustered Lasso was unsatisfying, they were able to use the estimated coefficients to order the covariates and run a successful Fused Lasso.

While some of these methods allow for the injection of *a priori* knowledge, they are not suited for complex interactions between the covariates, as it is the case between genes. Meanwhile, the grouping performed by the other methods can only be inferred from the data and do not necessarily reflect previously established interactions.

3.4.2 Graph based penalties

Biological interaction graphs are the result of many years of biological research. These graphs depict complex interactions between entities such as proteins or genes. Multiple propositions have been made to include these graphs in statistical models. Graph integration aims at smoothing the estimates over the network nodes and thus making the models more interpretable. Here we only focus on methods that add graph knowledge as a penalty to guide variable selection and some of their applications.

3.4.2.1 Graph penalty

Li and Li (2008) introduced a graphical penalty into their regression problem to smooth the estimated parameters over their graph. Instead of using an ElasticNet, the authors proposed to combine a Lasso penalty with a graph penalty that takes the form $\lambda_1||\boldsymbol{\beta}||_1 + \lambda_2||\boldsymbol{\beta}||^2_{\boldsymbol{L}}$ with $\boldsymbol{L}$ the normalised Laplacian associated to their graph. The authors demonstrated the grouping effect. In their study, the authors applied their model to glioblastoma data using genetic data (microarray). They showed that the performances of the graph constraint were similar to the ElasticNet while selecting more variables, which allowed them to exhibit sub-networks in the graph.

Grosenick et al. (2013) proposed another variant of the graphical constraint, which they called GraphNet. Instead of using the normalised Laplacian, they proposed using the Laplacian of the graph. Their aim was to obtain sparse but structured solutions by combining structured graph constraints with a global sparsity-inducing prior that automatically selects important variables. The authors applied their method to functional MRI data and a brain connectivity graph. They found the same grouping properties of the penalty as previously discussed.

GraphNet assumes that connected nodes in the graph should have similar coefficients, which is biased against negatively correlated variables. Du et al. (2016) proposes to alleviate this issue by using the absolute value of the coefficients. The authors combined their proposed penalty with a CCA model and called it the Absolute value GraphNet Structured CCA (AGN-SCCA). Applied to imaging-genetic data from an Alzheimer dataset (Alzheimer's Disease Neuroimaging Initiative (ADNI) database), this penalty outperformed similar methods in extracting the canonical correlations. Additionally, it yielded sparser results but identified relevant genetic and imaging markers.

Du et al. (2020) proposed two new graph-based penalties to capture better structures from the data. The first penalty, which they called Fused pairwise Group Lasso (FGL), takes the form $\lambda \sum_j \sqrt{\beta_j^2 + \beta_{j+1}^2}$. The second penalty, Graph guided pairwise Group Lasso (GGL), takes the form $\lambda \sum_{(j,k)\in\mathcal{V}} \sqrt{\beta_j^2 + \beta_k^2}$ with $\mathcal{V}$ the set of edges of the graph. Both penalties are not sensitive to the correlation sign. The authors combined the penalties with the CCA model and compared them with other penalties, including the AGN, on the ADNI dataset. They obtained better canonical correlations with FGL.

3.4.2.2 Applications

GraphNet has been applied using different statistical models in various biostatistical contexts. Watanabe et al. (2014) compared Fused Lasso, ElasticNet and GraphNet when applied to a Support Vector Machine (SVM) model. They tested the models on functional MRI data from a schizophrenia dataset (Center for Biomedical Research Excellence (COBRE)) and a spatial parcellation graph. Their results show that Fused Lasso and GraphNet both select spatially continuous regions, with the GraphNet selecting more variables. The increase in model complexity did not come at the cost of prediction accuracy. The classification accuracy with ElasticNet, GraphNet and Fused Lasso are comparable with similar sparsity levels.

Kim et al. (2020) studied the GraphNet penalty combined with a regression model and applied it to Parkinson's disease. They compared GraphNet with Group Lasso and Lasso. Their results show that the graph penalised model resulted in a more sparse model than the Group Lasso and a less sparse model than the Lasso. The GraphNet penalty model outperformed the two other models in predicting the severity of the disease.

Zhu et al. (2021) added the graph penalty to their Sparse Singular Value Decomposition (SSVD) model, which is called Sparse Network SVD (SNVD). The authors applied their method to integrate a prior gene interaction network from a protein-protein interaction network and gene expression data to identify underlying gene functional modules. Real cancer genomic data show that most co-expressed modules are significantly enriched on GO (Ashburner et al., 2000; Consortium, 2020)/KEGG (Kane-

hisa, 2000) pathways and correspond to dense sub-networks in the prior gene interaction network.



Chapter 4

Datasets Description and Preprocessing

DIPG pediatric gliomas are rare (about 40 new cases each year in France and 300 in Europe) which make clinical research study very difficult. Unfortunately, they are resistant to classical therapy and only transiently sensitive to radiotherapy. Recent efforts from the DIPG community established that 90% DIPG/DMG share a common molecular anomaly, namely a point mutation on Lysine 27 of the histone H3 regulatory tail (H3K27M). In the remaining cases, it was recently demonstrated that tumour cells overexpress EZHIP (Castel et al., 2020), leading to the same epigenetic dysregulation as H3K27M. Furthermore, the oncogenesis of these tumours seems to present general known mechanisms implicating the genes TP53, EGFR, PDGFRA, ACVR1, Src and mTOR pathway. The PHRC BIOMEDE and BIOMEDE 2.0 is an unprecedented effort in France and Europe that aims at collecting multi-omics and imaging data. The objectives are both clinical and fundamental: 1) to test new chemotherapies assigned based on patients' individual molecular profiles and to 2) to build a well characterized multi-modality measurement dataset to allow for a better understanding of the biological underpinnings of this tumour, to define new biomarkers, notably non-invasive ones based on imaging features, and derive some new hypnotical pharmaceutics in order to improve disease control.

During the duration of the PhD preparation, a significant fraction of the data could finally be accessed, but not the whole dataset. Without reconsidering the directions of the methodological work we chose to extend the datasets to tumours that have long been considered as a comparative context for DIPGs even though it is now known that these tumours are of different origin and prognosis. We have chosen high grade gliomas (glioblastoma) and low grade gliomas

Our work involves three different tumour types: Glioblastoma, LGG and DIPG. Glioblastoma and LGG were also chosen for their publicly available data as well as the multiple previous works about them. This allows us to benchmark our methods and compare our results with earlier findings. Finally we used the DIPG to apply our techniques and report the results. These results are given in terms of prediction or tumour characterisation.

This chapter describes the different datasets used with their various preprocessing steps. We start by defining the data sources and various acquisition parameters. Then we lay out the preprocessing pipeline for the radiomic feature extracting and the tumour segmentations. Finally, we describe the preprocessing applied to omic data.

Chapter Outline

Contents

4.1 Datasets Description 39
4.1.1 Glioblastoma 39
4.1.2 Lower Grade Gliomas 40
4.1.3 DIPG 41
4.1.4 Data availability 41
4.2 Preprocessing 42
4.2.1 Image preprocessing 42
4.2.2 Radiomic extraction 45
4.2.3 Mutation data preprocessing 46

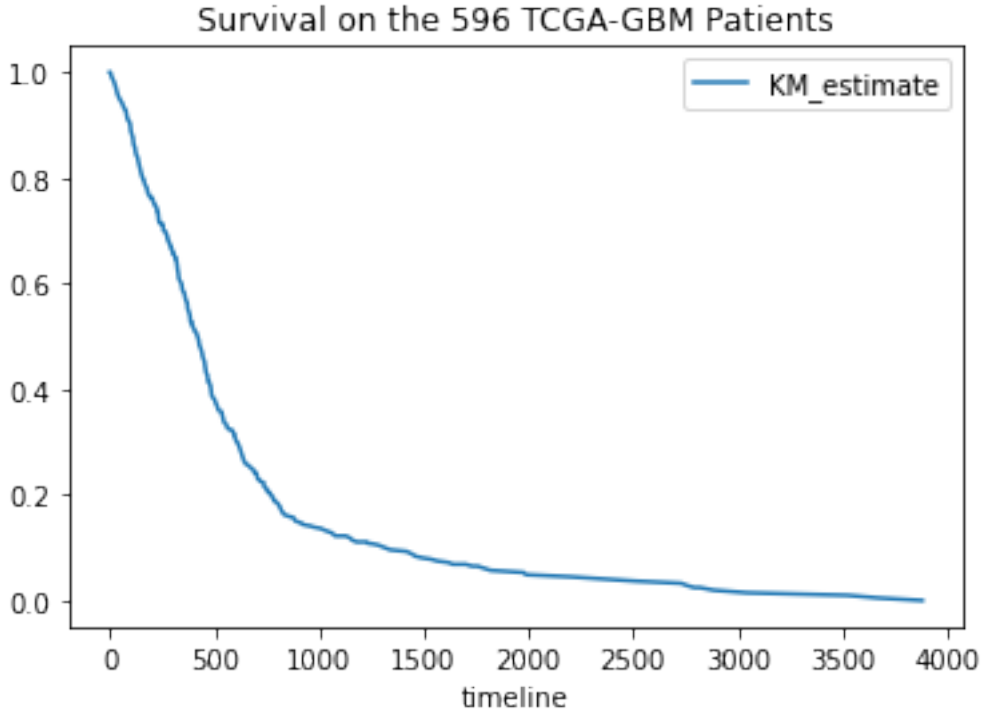


Figure 4.1: Kaplan Meier plot on the 561 TCGA-GBM patients with survival available. Duration is given in days

4.1 Datasets Description

4.1.1 Glioblastoma

To study glioblastoma, we used the TCGA-GBM dataset (Brennan et al., 2013; McLendon et al., 2008). The dataset comprises 602 patients from different sites, all diagnosed with glioblastoma. Data has been made available by The Cancer Genome Atlas (TCGA) Research Network: `https://www.cancer.gov/tcga`. Data were collected between 1989 and 2011. It includes Clinical data (602), Copy Number Variants (CNV) (599), micro RNA expression (miRNA) (5), RNA expression (RNAseq) (160), and somatic mutations (371). Clinical information, including all TCGA-GBM patient demographics, is summarised in Table 4.1, and the estimated survival probability using Kaplan–Meier in Figure 4.1.

We also used the public image dataset BraTS'19 (Bakas et al., 2017, 2019; Menze et al., 2015) (last access January 2020), which comprises 254 patients diagnosed for High Grade Gliomas. For each individual, 4 MRI volumes corresponding to T1w, CE T1w, T2w and FLAIR were available. These volumes were acquired with different clinical protocols and with various scanners from eight institutions and originated from different studies.

From the HGG set, we isolated the 102 patients diagnosed with Glioblastoma and belonging to the TCGA-GBM sub-cohort (and only kept 97 patients having manual segmentation provided). From now on, the HGG dataset refers to the $(254 - 102 = 152)$ patients from BRaTs'19 HGG deprived of TCGA-GBM. The HGG dataset will be divided into HGG^{train} and HGG^{val} and used for the training and validation sets for our segmentation method.

Table 4.1: TCGA-GBM Patient Demographics

Sex	n (% of total)
Male	366 (61)
Female	230 (38)
Total	602
Age	Average (range), in years
	58.36 (11 - 89)
Survival	Average (std), in days
Days to death	476.28 (529.01)
Days to last follow-up	415.34 (482.37)
Non-censored/Censored	447/149

Table 4.2: TCGA-LGG Patient Demographics

Sex	n (% of total)
Male	285 (55)
Female	230 (45)
Total	516
Age	Average (range), in years
	43.46 (14 - 87)
Survival	Average (std), in days
Days to death	1205.80 (1121.42)
Days to last follow-up	604.62 (846.96)
Non-censored/Censored	92/422

4.1.2 Lower Grade Gliomas

To study Lower-Grade Gliomas (LGG), we worked with the TCGA-LGG dataset (TCGA, 2015) obtained from The Cancer Genome Atlas (TCGA) project (`http://cancergenome.nih.gov`). We used the data as they have been made available on `https://www.openml.org/` by Herrmann et al. (2020). Data was collected between 1992 and 2013 and comprises five groups of variables (five blocks), including Clinical records, Copy Number Variants (CNV), micro RNA expression (miRNA), gene expression (RNAseq), and somatic mutations, obtained for 419 patients. Clinical information, including all TCGA-LGG patient demographics, is summarised in Table 4.2, and the estimated survival probability using Kaplan–Meier in Figure 4.2.

Additionally, we used the MR scans of 108 patients obtained from the BRaTs'19 dataset, 76 of which are publicly available. We used the 76 initially obtained images to assess our segmentation methods and all 108 images for radiomic extraction and data integration. All four MRI modalities corresponding to T1w, CE T1w, T2w and FLAIR were available for each individual. Similarly to TCGA-GBM, these volumes were acquired from five different institutions using various clinical protocols and scanners.

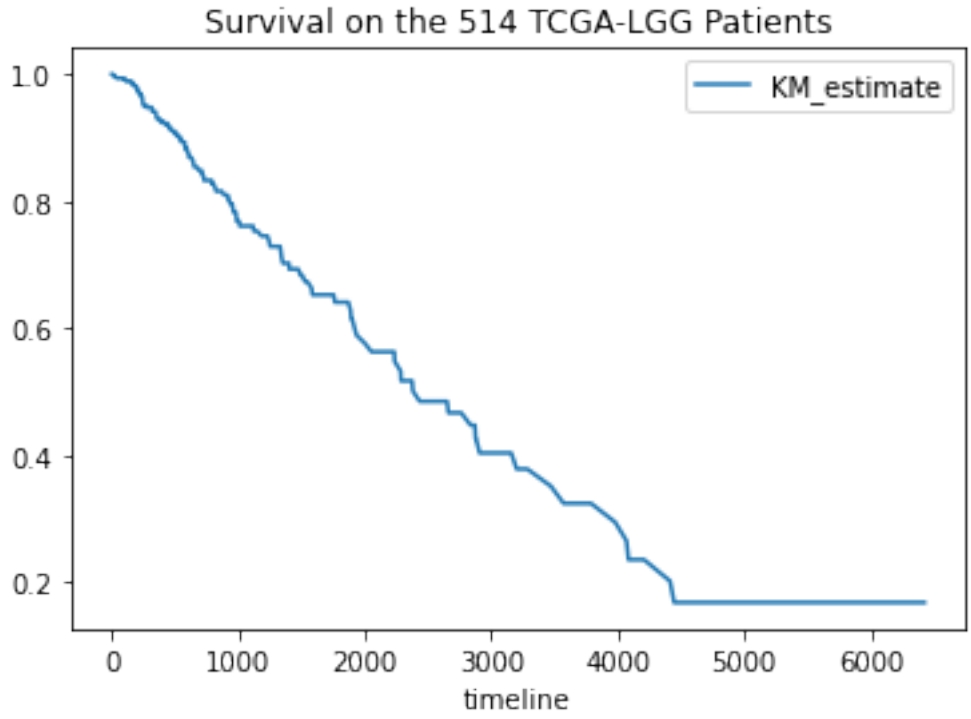


Figure 4.2: Kaplan Meier plot on the 561 TCGA-LGG patients with survival available. Duration is given in days.

4.1.3 DIPG

To study the DIPG, we used two similar but distinct cohorts. The PREBIOMEDE cohort includes pediatric patients referred to the *hôpital Necker enfants malades*, Paris, France, between 2011 and 2013. The inclusion criteria were based on the combined information of the DIPG diagnosis by an MR scan and a stereotactic biopsy. The dataset includes MR scans of 30 patients across 71 sessions. All the cases have the T2w and FLAIR modalities available. Manual delineation of the whole tumour was achieved for all test cases and checked on the most relevant modality (for some patients, ASL and diffusion MR images were used) (Calmon et al., 2017, 2021). These segmentations were obtained on images with a 4 mm^3 resolution. Informed consent of the use of clinical and radiological data was obtained and the protocol was approved by the local ethics committee of the reference institution.

The BIOMEDE project is an initiative led by the French Society of Paediatric Oncology and the European consortium "Innovative Therapies in Children with Cancer (ITCC)". The cohort includes 189 patients diagnosed with DIPG from multiple European institutions. Available data includes MR scans, gene expression and somatic mutation profiles. No delineation of the tumours is available.

The PREBIOMEDE cohort was only used to assess our segmentation method. *As of 09/12/2022, no clinical data was accessible for the BIOMEDE patients.*

4.1.4 Data availability

The complete data availability description can be found in Table 4.3. We used the TCGA-GBM, TCGA-LGG and PREBIOMEDE to assess the performance of our segmentation model (see chapter 5). Since the mutation and image require prior preprocessing, we only used the CNV, miRNA and

RNAseq data to study variable selection in a multi-block framework from the TCGA-LGG dataset (see chapter 6). We used RNAseq, mutations and images for the radio-genomics integration (see chapter 7). Survival data were unavailable for BIOMEDE, and TCGA-GBM did not have sufficient data which precluded their use in the latter study.

Table 4.3: Number of samples available of the different datasets used

	TCGA-GBM	TCGA-LGG	BIOMEDE
# Images	102	108	89
# Clinical	602	516	
# CNV	599	515	
# miRNA	5	514	
# RNAseq	160	512	69
# Mutations	371	510	87
# CNV & miRNA & RNAseq	5	419	
# RNAseq & Mutations & Images	20	90	28

4.2 Preprocessing

To use the different data at hand, each block underwent preprocessing steps. This section details the preprocessing pipeline applied to each block.

4.2.1 Image preprocessing

Radiomic analysis utilises intensity distributions to extract features. In order to obtain comparable intensity distributions within the studied cohorts, several preprocessing steps have been applied to each image. First, different MRI sequences of the same patient were skull-stripped and co-registered to each other. Then, image intensities were bias field corrected and normalised into a common distribution, allowing comparison across different patients. Finally, multiple radiomic features were automatically calculated using previously extracted regions of interest (ROI).

For the segmentation, we used the bias field corrected images. Then value harmonisation was obtained with a min-max intensity normalisation. Our choice of normalisation is justified by our intent to enhance tumour intensities, thus making our detection and segmentation tasks easier. The images were min-max normalised using the 5% and 95% percentiles to discard outliers. Out-of-range values were capped as in Eq 4.1 where v and $\hat{v}$ are respectively the original and normalised grey level of a generic voxel of the image $\mathbf{v}$:

$$\hat{v} = max\left(min\left(\frac{v - percentile(\mathbf{v}, .05)}{percentile(\mathbf{v}, .95) - percentile(\mathbf{v}, .05)}, 1\right), 0\right) \tag{4.1}$$

Figure 4.3 illustrates the preprocessing steps.

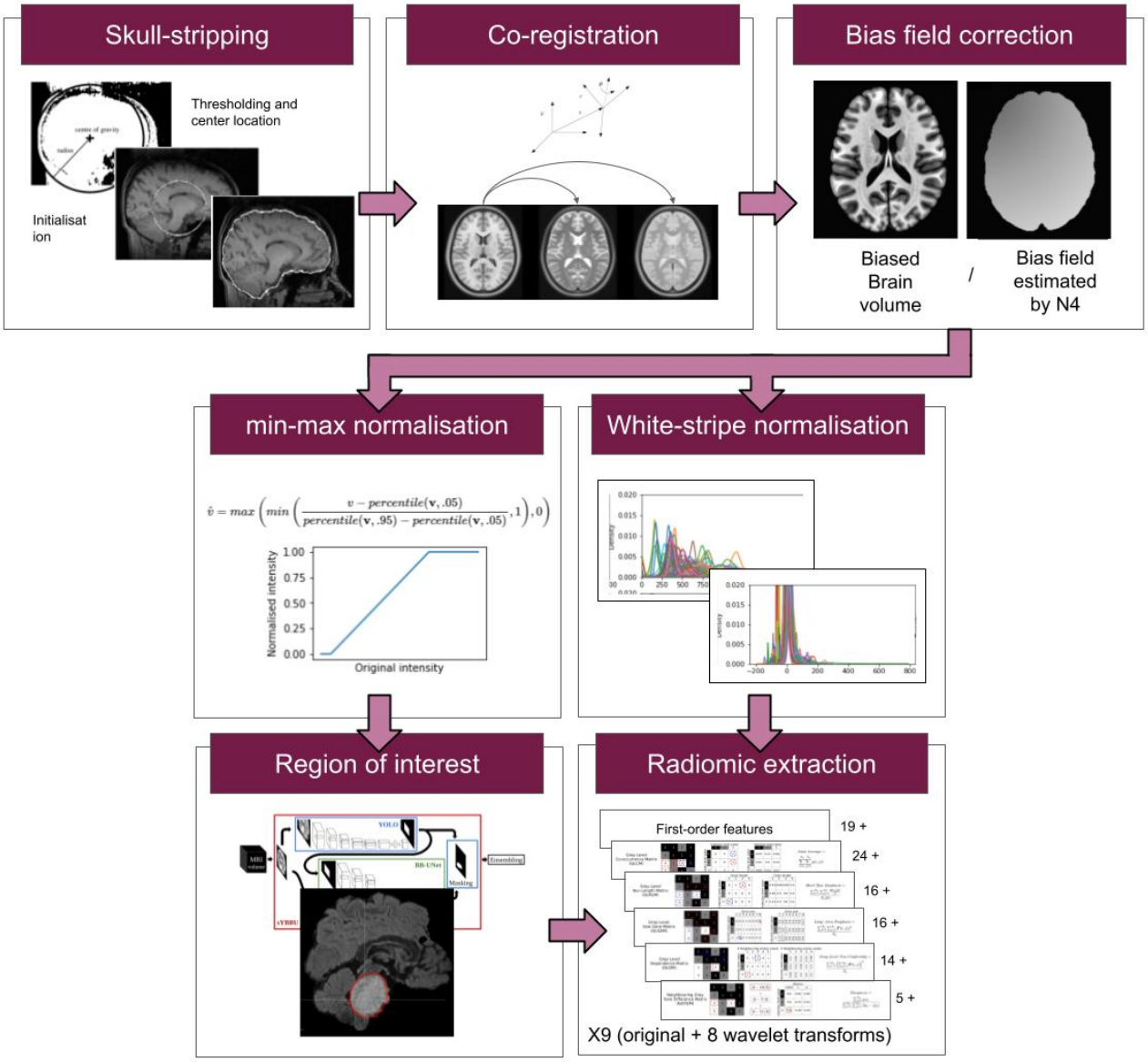


Figure 4.3: The different pre-processing steps applied to MR images

4.2.1.1 Skull-stripping

Skull-stripping aims at finding a 3D brain mask in the MR scan, which removes the skull. We used the FSL BET (Smith, 2002) software. The algorithm first estimates the brain volume using the intensity distribution. This is done by removing outlier intensities from the image (2% and 98% percentile), then binarising the image using a threshold that lier at 10% of the intensity range. This allows the algorithm to estimate the brain centre of gravity (CoG) and its spherical radius. Starting from a spherical tessellated surface centred on the CoG with half the radius estimated, the surface iteratively expands until it reaches a stationary state.

4.2.1.2 Co-registeration

The co-registration problem aims at aligning a moving image on a reference volume. It is done by finding the best geometric transformation that maximizes the similarity between the transformed moving image and the reference volume. We used rigid co-registration to align each patient's different scans to their reference modality, which defines a transformation with six degrees of freedom (three translations and three rotations). In our case, the reference modality is the T1w, which is often the highest-quality image in terms of resolution and contrast.

We used the FSL FLIRT (Jenkinson and Smith, 2001; Jenkinson et al., 2002) for the co-registration. The software utilises a hybrid local-global optimisation algorithm on the same image at multiple resolutions. We used the normalised mutual information as our cost function, defined as $C(\boldsymbol{X}, \boldsymbol{Y}) = \frac{H(\boldsymbol{X},\boldsymbol{Y})}{H(\boldsymbol{X})+H(\boldsymbol{Y})}$ where H is the standard entropy and $\boldsymbol{X}$ and $\boldsymbol{Y}$ are two volumes with the same dimensions.

4.2.1.3 Bias field correction

Bias field correction aims at removing unwanted low-frequency intensity nonuniformity from the images. We used the N4ITK algorithm (Tustison et al., 2010). As the N3 bias correction algorithm (Sled et al., 1998), the bias field is modelled with a multiplicative effect (i.e. $\boldsymbol{I}^{biased}(x) = \boldsymbol{I}^{unbiased}(x)\boldsymbol{F}(x)$ with I the image and F the field at a voxel x). Their approach consists of finding a smooth, slowly varying gaussian field that maximises the frequency of the unbiased image. N4 improved the optimisation routine and the B-spline approximation method over its predecessor. The implementation used was provided by Advanced Normalisation Tools (ANTs) Software `https://github.com/ANTsX/ANTs`.

4.2.1.4 White-stripe normalisation

Intensity normalisation aims at reducing discrepancies between intensity distributions within tissue classes (white matter, grey matter ...) across the subjects. We used the white stripe normalisation (Shinohara et al., 2014). The method focuses on matching the distributions between the different patients on the white matter. The intensity distribution in other tissues is adjusted accordingly:

$$\boldsymbol{I}^{normalised} = \frac{\boldsymbol{I} - \mu^{WM}}{\sigma^{WM}} \tag{4.2}$$

With $\boldsymbol{I}$ the 3D image. To estimate the mean (μ^{WM}) and standard deviation (σ^{WM}), the distribution of intensities within a 4 cm rectangle at the centre of the head is considered. This rectangle

contains a large amount of white-matter voxels. The distribution is smoothed using B-splines of order four. The μ^{WM} corresponds to the last peak in the smoothed distribution for the T1w modality and the highest peak for the T2w and FLAIR modalities. These intensities correspond to estimated white-matter voxels. To estimate σ^{WM}, the set of voxels having intensities within the range $f^{-1}(f(\mu^{WM})-\tau)$ and $f^{-1}(f(\mu^{WM})+\tau)$ are considered. With f the smoothed estimated intensity distribution and τ is a quantile tolerance, which takes the value 0.05 as recommended by the original paper.

We used the hybrid white-stripe to normalise the image, which takes the intersection of estimated white-matter voxels from multiple modalities of the same patient. We used in-house implementation of white-stripe, using python language (Van Rossum and Drake, 2009). Figure 4.4 show the effect of the white-stripe normalisation on the intensity distribution of LGG MR images.

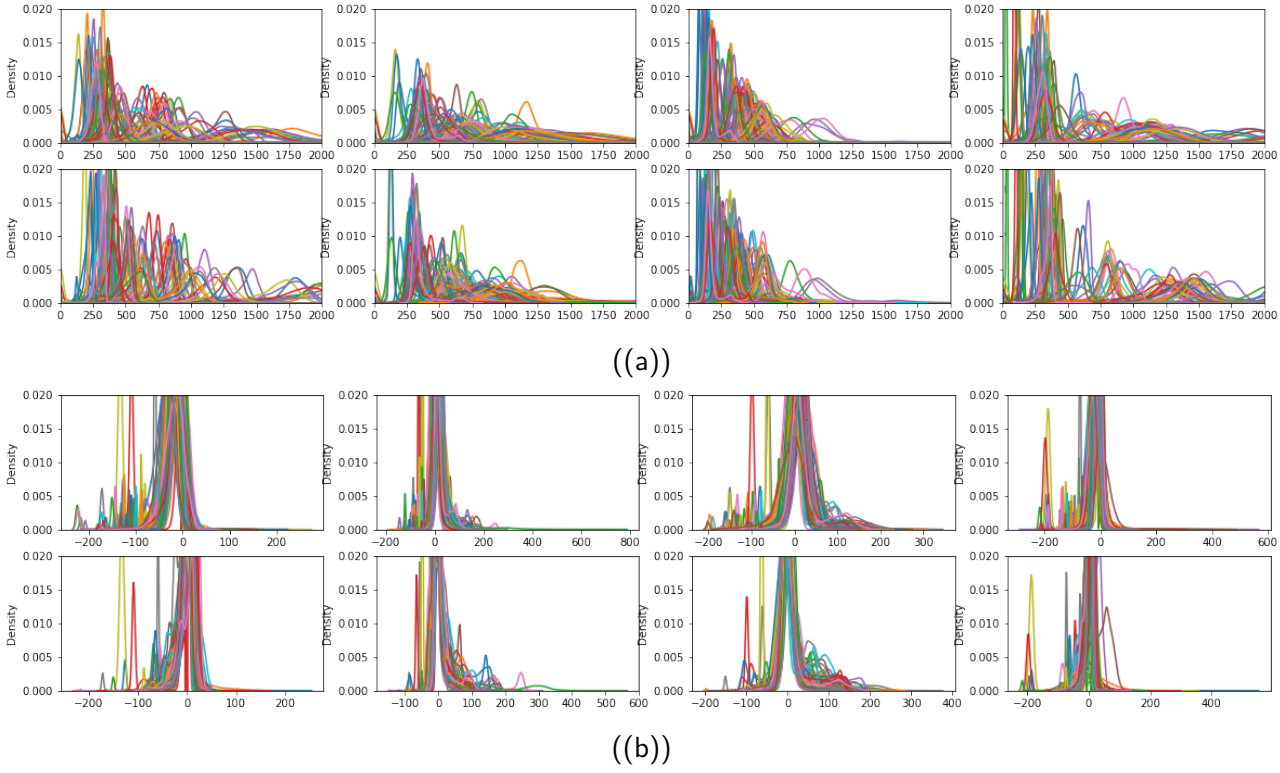


Figure 4.4: Intensity distribution on the MRI image for each patient on the LGG dataset. Columns correspond to the modalities T1w, T2w, FLAIR, CE T1w. First row represent distribution of the white matter voxels, and second row for the grey matter. (a) Before white-stripe and (b) after white-stripe normalisation

4.2.2 Radiomic extraction

We extracted features from the original images and eight wavelet-transformed images. This resulted in 846 ($(19 + 24 + 16 + 16 + 14 + 5) \times 9$ (original + 8 wavelet transformed images)) radiomic features extracted from each modality. Note that we did not use shape-related features. We used the pyradiomics software (van Griethuysen et al., 2017) to extract the radiomic features.

4.2.3 Mutation data preprocessing

Statistical analysis of somatic mutations is challenging. Most tumour mutations are only present in a small sample of patients. The mutated gene may or may not be impactful on the pathology. Statistically, deciding the relevance of these genes from the small observed samples is infeasible. Additionally, mutations in genes may lead to different results. Some mutations lead to the dysfunction of the gene, which promotes the cancerous behaviour of the cells. Other mutations on some genes may impact other genes in the same biological pathway, even if the latter are not mutated. Finally, other mutations may not cause tumoural cells but rather be a consequence of the rapid cell division typical of cancer.

Starting from the matrix of somatic mutations (samples × genes), we used network smoothing proposed by Vanunu et al. (2010) which diffuses the mutation information in the matrix according to the network. This procedure normalise the somatic mutation profiles under the hypothesis that mutation or dysfunction in a gene leads to the disturbance of the behaviour of other genes it interacts with. Using a gene-gene interaction graph, the mutation propagation is given by the equation:

$$\boldsymbol{X}_{t+1} = \alpha \boldsymbol{X}_t \boldsymbol{D}^{-\frac{1}{2}} \boldsymbol{A} \boldsymbol{D}^{-\frac{1}{2}} + (1-\alpha)\boldsymbol{X}_0 \tag{4.3}$$

Where $\boldsymbol{X}_t$ is the somatic mutation profile at iteration t, with $\boldsymbol{X}_0$ the initial somatic profiles. $\boldsymbol{D}$ is the diagonal matrix of the gene degrees (number of neighbours in the graph). $\boldsymbol{A}$ is the binary adjacency matrix indicating if two genes are neighbours in the graph. α is a tuning parameter between 0 and 1, governing the distance that a mutation signal is allowed to diffuse through the network during propagation. We computed the propagation t times until convergence was found for the Frobenius norm.

Gene-gene interaction graphs usually describe complex interactions between genes. Suppose the case where a gene implicated in multiple biological pathways is mutated. We hypothesise that the influence on each of the biological pathways (and on each gene involved in these pathways) is lesser compared to a gene implicated in fewer pathways. The distance between two genes in the graph does not reflect their influence on each other. To solve this issue, we used the influence graph computed from the gene-gene interaction graph proposed by Vandin et al. (2011b). The method uses the heat-diffusion model on the gene-gene interaction graph to compute the influence. Let $\boldsymbol{L}_\gamma = \boldsymbol{L} + \gamma I$, where L is the Laplacian of the interaction graph, and γ represents the diffusion rate. The influence scores are given by $\boldsymbol{L}' = \boldsymbol{L}_\gamma^{-1}$ and $L'_{i,j}$ gives the influence of gene i on the gene j. The influence graph is defined as $G = (E, V)$, where E is the set of genes, and for each gene, its neighbours are the n genes with the highest influence score on it (by default $n = 10$). Note that the graph is undirected.

In our proposed thesis, we decided to preprocess the somatic mutation profiles using both the original Pathway Commons gene-gene interaction and the influence graph derived from it. We used $\gamma = 10^{-2}$ to compute the influence graph. For the propagation algorithm, we set $\alpha = 0.5$, as recommended for Pathway Commons by Hofree et al. (2013).

* * *
* *
*

Part II

Contributions

Chapter 5

Brain Tumour Segmentation

Tumour lesion segmentation is a key step to study and characterise cancer from MR neuroradiological images. Nowadays, numerous deep learning segmentation architectures have been shown to perform well on the specific tumour type they are trained on (e.g. glioblastoma in brain hemispheres). But, a high performing network heavily trained on a given tumour type may perform poorly on a rare tumour type for which no labelled cases allows training or transfer learning. Yet, because some visual similarities exist nevertheless between common and rare tumours, in the lesion and around it, one may split the problem into two steps: object detection and segmentation. For each step, trained networks on common lesions could be used on rare ones following a domain adaptation scheme without extra fine-tuning.

This study evaluates the impact of adding an object detection framework into brain tumour segmentation models, especially when the models are applied to different domains. We identify object detection as a simpler problem that can be injected into a segmentation model as an *a priori*, and which can increase the performance of the segmentation models. All models were trained on glioblastoma, and they were evaluated on their performances on the glioblastoma, LGG and DIPG. This work has been the subject of a journal article (Chegraoui et al., 2021a).

Availability: Source code is freely available at `https://github.com/neurospin-projects/2021_hchegraoui_DetectSegmBIOMEDE/tree/main`

Chapter Outline

Contents

5.1 Introduction . . . 51
5.2 Detection-segmentation combination strategy . . . 52
5.2.1 Combination strategies . . . 52
5.2.2 Final masking . . . 52
5.2.3 Input data . . . 53
5.2.4 Ensembling the inferences . . . 54
5.3 Reusing off-the-shelf networks . . . 54
5.3.1 You Only Look Once (YOLO) . . . 54
5.3.2 UNet and BB-UNet models . . . 56
5.3.3 Deepmedic . . . 56
5.4 Experimental designs . . . 57
5.5 Benchmark results . . . 58
5.5.1 Object-detection results . . . 58
5.5.2 Segmentation results on DIPG . . . 63
5.6 Discussion . . . 65
5.7 Conclusion . . . 67

5.1 Introduction

Learning from images classically requires large cohorts for which tumours are finely delineated. The rarity of DIPG added to the fact that segmentation is not part of the clinical routine procedure makes it difficult to obtain robust statistical classifiers or predictors. Automatic tumour segmentation based on transfer learning could theoretically alleviate this problem, circumventing the small number of manual delineations which prevent directly and efficiently training these segmentation models.

In general, DIPG tumours have a central location involving more than 50% of the pons (Warren, 2012). On MRI scans and due to its infiltrating nature, the tumour appears as an intrinsic expansion of the brainstem and not as a distinct foreign mass compressing the pons. However, the tumour is not always restricted to the pons and it can infiltrate other compartments of the central nervous system such as the cerebral peduncles and supratentorial midline or the cerebellum (Hankinson et al., 2011). The deformation of the pons induced by the tumour and its infiltrating nature makes its detection and delineation non-trivial. On the T2w scans, the tumour presents a hyper-intense signal while it appears hypo-intense with indistinct tumour margins on T1w scans. Enhancement following gadolinium injection (T1Gd) is inconstant and often absent. Finally, the tumour is relatively homogeneous on FLAIR modality (Warren, 2012). Figure 5.1 exhibits a DIPG tumour on different modalities.

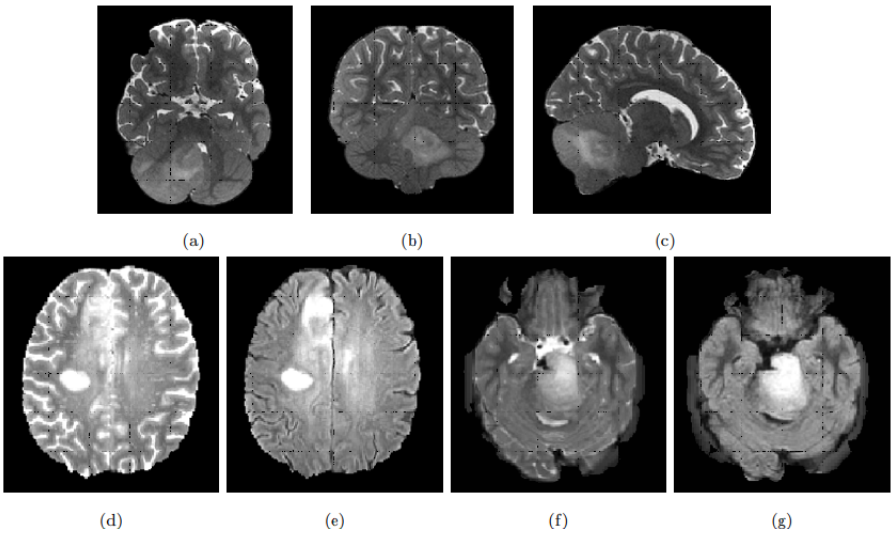


Figure 5.1: Top row: Example of T2w MRI scan of a patient with DIPG tumour extending beyond the pons. (a) Axial, (b) Coronal and (c) Sagittal slices. Bottom row: Example of glioblastoma and DIPG MRI scans, in axial slices. (d) glioblastoma T2w, (e) glioblastoma FLAIR, (f) DIPG T2w and (g) DIPG FLAIR. Tumours appear on the images as hyper-signal.

DIPG shares some of its visual characteristics with the glioblastoma, especially on T2w and FLAIR modalities. However, glioblastoma presentation differs on T1w and T1Gd, with the absence of a necrosis component in the DIPG, and the gadolinium enhancement which is more intense and always present for the glioblastoma (Villanueva-Meyer et al., 2017). Our aim is to exploit the existing macroscopic visual similarities of DIPG with glioblastoma or low-grade gliomas, to train a two-step robust model able to infer DIPG segmentations.

Because rare tumours present some visual similarities with common tumours, in the lesion and

around it, one may split the tumour delineation problem into two steps: object detection and segmentation. For each step, trained networks on common lesions could be used on rare ones following a domain adaptation scheme without extra fine-tuning. Our work suggests different architectures to solve the segmentation and detection tasks and two combination strategies of the two tasks. We assessed the performances of our strategies in three different configurations: i) using the same tumour as in training, ii) using a different tumour located in the supratentorial region as in training and finally iii) using a different rare tumour located in the brainstem, an unseen region during the training.

5.2 Detection-segmentation combination strategy

5.2.1 Combination strategies

In order to obtain a robust segmentation of brain tumours, we combined proven object-detection models and segmentation models. Considering we could not directly learn from a few labelled DIPG examples, we decided to train our models with HGG^{train} examples. The HGG and DIPG tumours present both similarities and differences (see Figure 5.1). Tumour intensities have comparable characteristics, while the ages of the patients, tumour locations and image qualities differ. We hypothesised that in a restrained zone around the tumour, HGG and DIPG present enough visual similarities to allow the training of a segmentation model from the sole HGG data and which would be able to segment both types of tumours reliably. We used an object detection model to define these restrained zones around the tumour and bypass the dissimilarities between the two cohorts.

We chose You Only Look Once (YOLO) (Jocher et al., 2020; Redmon et al., 2016) as our object-detection framework. For the segmentation, we benchmarked UNet (Ronneberger et al., 2015) and Bounding-Box UNet (BB-UNet) (Rosana et al., 2020). It must be noted that both UNet and BB-UNet receive the whole 2D images as input for the training and inference; additionally, BB-UNet receives also an *a priori* bounding-box used internally to (non-exclusively) focus the learning segmentation process. Consequently, we examined two different procedures for combining the object-detection and segmentation. We called our first procedure Parallel YOLO UNet (pYU). In pYU, both YOLO and UNet are trained independently. In the inference phase, YOLO-generated bounding-boxes are merely used to mask UNet predictions, thus eliminating all segmented voxels outside the bounding-box. In our second approach, called Sequential YOLO BB-UNet (sYBBU), YOLO and BB-UNet are trained independently, but during inference, YOLO-generated bounding-boxes are provided as additional input to BB-UNet. Similarly to pYU, we used the bounding-boxes to mask the segmentation output. Figure 5.2 illustrates the two approaches, pYU and sYBBU, sketched in their inference stage.

5.2.2 Final masking

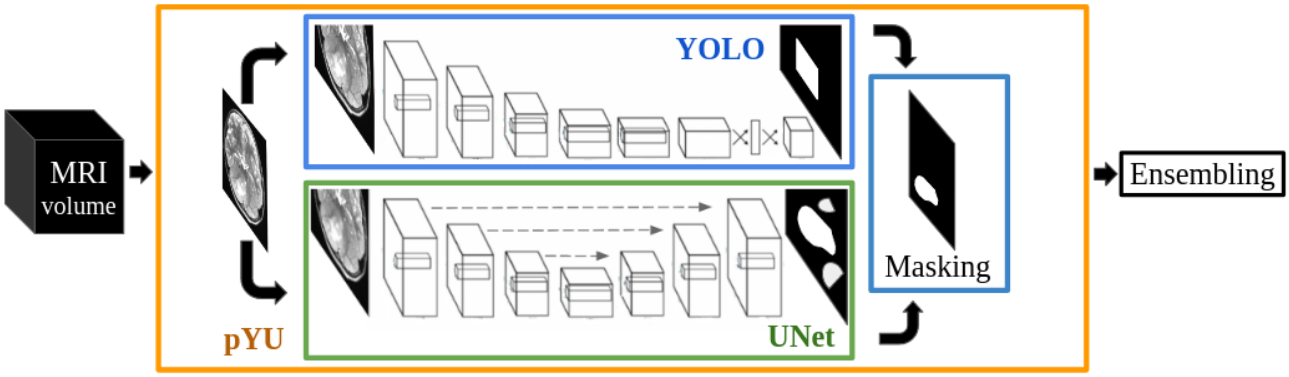
Let SEG, $BBOX$ and GT be the sets of voxels belonging to the predicted segmentation, predicted bounding-box and ground truth respectively. Our approaches introduce a masking phase. Masking (versus no masking) will affect precision/recall scores in an anticipated direction, if we make an assumption that will be checked in our results: let v be a voxel and P a probability:

$$P(v \in GT | v \in (SEG \cap BBOX)) \geq P(v \in GT | v \in (SEG \setminus BBOX)) \tag{5.1}$$

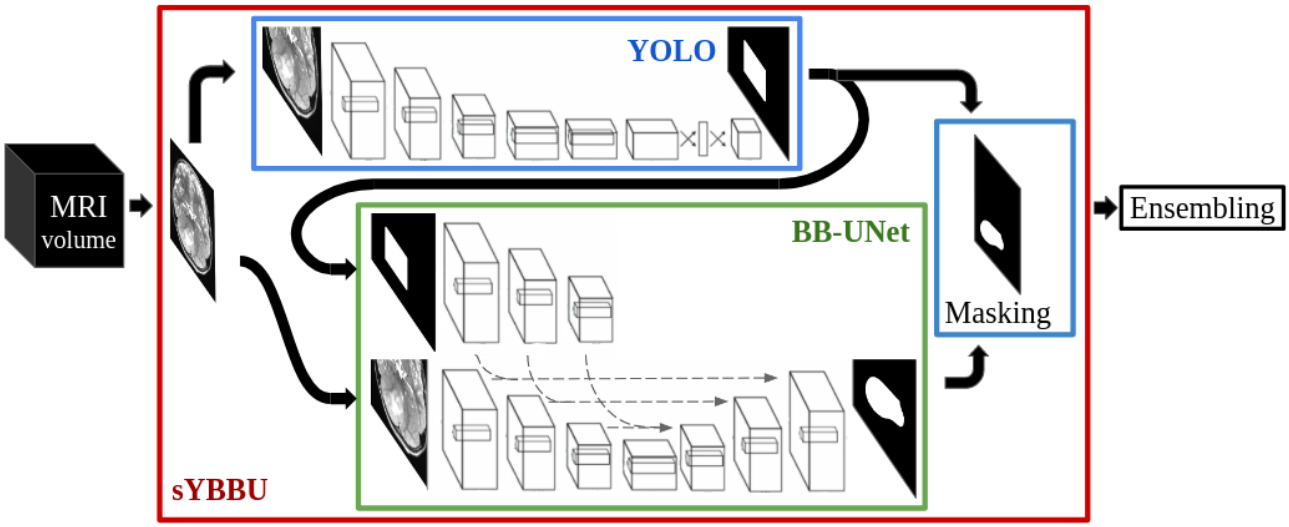
Under this assumption, it is more likely to find a true positive inside the bounding-box than outside. Thus, it follows that:

$$\begin{aligned} Precision(SEG \cap BBOX) &\geq Precision(SEG) \\ Recall(SEG \cap BBOX) &\leq Recall(SEG) \end{aligned} \tag{5.2}$$

Masking will always result in a decrease in the recall (amount of the tumour detected). However, we assume that the decrease in the recall will be outweighed by the increase in precision.



((a)) Parallel YOLO UNet (pYU) model



((b)) Sequential YOLO BB-UNet (sYBBU) model

Figure 5.2: The two approaches pYU and sYBBU

5.2.3 Input data

We decided to train all our models on 2D slices as we can extract a greater number of different training examples from each 3D volume. Furthermore, we trained all our models only on axial slices since these slices have the best resolution in all the studied cohorts.

In this work, we only used the T2w and the FLAIR modalities. Our choice is justified by several considerations. First, HGG and DIPG MRI scans present similar local tumour patterns mostly in these

two sequences. Furthermore, we are only interested in a binary segmentation, and the modalities which best reveals all the different compartments of the tumour are the T2w and the FLAIR. Finally, the DIPG is representative of DIPG data obtained in clinical context, which contains many subjects with missing sequences, but most subjects have either the T2w or the FLAIR available.

5.2.4 Ensembling the inferences

Each model is trained twice : i) on the T2w and ii) on the FLAIR modalities. This makes our approach resilient to missing data. Depending on the data at hand, we retain the inference obtained from the single modality available or we combine the two inferences. In this latter case, we merged the predictions using a weighted average. As described by (Shahhosseini et al., 2020), we propose to find the optimal weights ω^* according to the optimisation problem, where N is the number of validation cases, $\hat{y}_v^{FLAIR}$ (resp. $\hat{y}_v^{T2w}$) is the model's confidence scores for the FLAIR (resp. T2w) on a voxel, $label(\hat{y}_v)$ is the thresholded confidence score and takes the values 0 (for non-tumoural voxels) or 1 (for tumoural voxels) :

$$\begin{gathered} \text{with } \hat{y}_v(\omega) = \omega \times \hat{y}_v^{FLAIR} + (1-\omega) \times \hat{y}_v^{T2w} \\ \omega^* = \underset{\omega}{argmin} \left(-\frac{1}{N} \sum_{k=1}^{N} \left(\sum_{v \in DISCR} GT_v \times \log(\hat{y_v}(\omega)) + (1 - GT_v) \times \log(1 - \hat{y_v}(\omega)) \right) \right) \\ \text{s.t. } 0 \leq \omega \leq 1 \end{gathered} \tag{5.3}$$

To find the optimal weight, we used the HGG^{val} set. To better fit our needs, we adapted the optimisation problem by only considering the $DISCR = \{v | label(\hat{y}_v^{FLAIR}) \neq label(\hat{y}_v^{T2w})\}$ (i.e. the set of voxels where T2w output label and FLAIR output label are different). This is justified by the fact that the weighted average does not change the predicted label if both models predict the same label.

Ensembling was used both after the detection inferences and after the segmentation inferences. More precisely, we used the confidence scores of the YOLO models to compute an ensembled bounding-box prediction, and the UNet and BB-UNet scores for an ensemble segmentation prediction. For the segmentation, the ensembling phase is prior to the bounding-box masking.

We used Brent's method (Brent, 1973) to solve the optimisation problem (Eq. 5.3) and SciPy's optimisation package (Virtanen et al., 2020).

5.3 Reusing off-the-shelf networks

5.3.1 You Only Look Once (YOLO)

You Only Look Once (YOLO) framework is a multi-scale object detection neural network. YOLO is designed to detect multiple objects of different classes on natural images. The input image is divided into an $S \times S$ grid of cells. The cell where the centre of the object falls into is responsible for predicting the bounding-box and class of the objects. Each cell predicts B bounding-boxes, a confidence related to the existence of an object in each bounding-box, and conditional class probabilities related to the object instances.

We used the YOLOv5, implemented by Ultralytics (Jocher et al., 2021). This network was pre-trained on 416×416 images from the Common Objects In Context (COCO) dataset (Lin et al., 2014). We used transfer-learning to fine-tune the model parameters for the tumour detection task. We resized our 250×250 input images to the 416×416 dimension using zero padding. The model was trained to detect the smallest bounding-box around the tumour. Since YOLO is pre-trained on RGB images, we transformed our grey-scale images into RGB images by copying our input image into the three channels.

For the hyper-parameter B, we kept its default value 3. Because YOLO makes predictions with a multi-scale approach, S took successively 3 values in which the prediction is made. For 416×416 input images, the values taken were 13, 26, 52. Furthermore, we were interested in detecting only one object, namely the tumour (using the –config file requested by the software).

Finally, starting from the pretrained model, we fine-tuned it for 100 epochs, using an initial learning rate of 0.001. Other default training parameters were kept. Main parameters are listed in the Table 5.1

Table 5.1: YOLO parameters used for training

		Parameters
Gradient	lro	0.00320
	lrf	0.12000
	momentum	0.84300
Training	weight_decay	0.00036
	warmup_epochs	2.00000
	warmup_momentum	0.50000
	warmup_bias_lr	0.05000
Loss function	box	0.02960
	cls	0.24300
	cls_pw	0.63100
	obj	0.30100
	obj_pw	0.91100
	iou_t	0.20000
	anchor_t	2.91000
Augmentation	degrees	0.37300
	translate	0.24500
	scale	0.89800
	shear	0.60200
	perspective	0.00000
	flipud	0.00856
	fliplr	0.50000

We assume that tumours are 3D-connected-component volumes. However, as the detection model took axial 2D images as input, there was thus no guarantee to obtain a connected component object in a plane perpendicular to the axial plane. The model might miss the tumour on some slices of the volume, or detect tumours on isolated slices. To overcome this issue, we used a morphological closing of the bounding-boxes, followed by an opening, along the perpendicular axis to the axial plane, with a kernel size of (1,1,6) voxels.

5.3.2 UNet and BB-UNet models

For the segmentation step of the tumours, we studied two models. Firstly, we used UNet, a fully convolutional neural network, which is classically used for biomedical image segmentation (Ronneberger et al., 2015). Similar to an auto-encoder, it has two paths: an encoding path consisting of the stacking of convolutions, non-linear activations and max-pooling; and a decoding path which consists of convolutions, non-linear activations and transposed convolutions. Skip connections are used between each encoding layer and its symmetric decoding layer. Our network is 5 levels deep on each path. We used the rectified linear unit (ReLU), defined as $ReLu(u) = max(u, 0)$, as the non-linear activation function.

Secondly, we used the BB-UNet (Rosana et al., 2020) model, whose architecture is similar to that of the UNet, except that it takes a binary bounding-box mask as additional input. We added to the UNet a bounding-box path parallel to the encoder path. The binary mask follows similar transformations to the main image. At each skip connection, we carry out an element-wise multiplication between the encoded image and encoded bounding-box. The role of these bounding-boxes is to discourage the network to look beyond it. Our BB-UNet models were trained using the ground truth bounding-boxes obtained as the smallest bounding-boxes comprising all the tumour mask (in 2D). As stated before, we used YOLO-predicted bounding-box during the inference phase.

For both models, the last layer has a soft-max activation, and we used a binary cross-entropy as a loss function, defined as :

$$Loss(\hat{\boldsymbol{y}}, \boldsymbol{y}) = -\boldsymbol{y} \cdot \log \hat{\boldsymbol{y}} + (1 - \boldsymbol{y}) \cdot \log (1 - \hat{\boldsymbol{y}}) \tag{5.4}$$

where $\boldsymbol{y}$ an $\hat{\boldsymbol{y}}$ are the ground-truth labels and the network confidence score matrices respectively.

We implemented both UNet and BB-UNet using Pytorch (Paszke et al., 2019). The neural networks were trained on 250×250 grey-scale images for 100 epochs, with an initial learning rate of 0.001 and the Adam optimiser (Kingma and Ba, 2015). Given the size of the used dataset (≥ 9000), no data augmentation was used.

5.3.3 Deepmedic

We compared all our results to a reference, patch-based brain lesion segmentation network, namely Deepmedic (Kamnitsas et al., 2017). Deepmedic is an 11-layers deep, double-pathway, multi-scale, 3D CNN. Deepmedic achieved state-of-the-art results on brain tumour segmentation on BraTS'15, and it is continuously updated. We trained Deepmedic on mono-channel twice, with T2w and FLAIR, to make the results comparable with our approaches. Contrary to our models, input images for Deepmedic were normalised using z-scores to remain in line with the network procedure.

We used the implementation from `https://github.com/deepmedic/deepmedic`. We kept the default values for the hyperparameters as proposed by the original paper, including the number of layers, filters, learning rate, and optimiser.

5.4 Experimental designs

To conduct our experiments, we divided the HGG dataset into 90% training set and 10% validation set. We tested all the models on the TCGA-GBM dataset, the LGG dataset and 30 patients (71 sessions) of DIPG. Table 5.2 sums up the dataset sizes.

Table 5.2: Dataset sizes of the different training and testing sets

dataset	HGG^{train}	HGG^{val}	TCGA-GBM	LGG	DIPG
number of patients	142	15	97	76	30 (71 sessions)
number of 2D images	9533	783			

To assess the performance of our approaches on the different test datasets, we used the provided segmentation labels to compute precision and recall, alongside the Dice index. These metrics were measured after 3D reconstruction of the binary masks. We note that the object-detection outputs are also binary masks.

On the TCGA-GBM dataset, we performed a correlation analysis between the ensembled bounding-box performance and the ensembled segmentation performance, in order to establish the impact of the object-detection step on the final segmentation. Furthermore, since BB-UNet models were trained with the ground-truth bounding-boxes while inference was performed using YOLO predicted bounding-boxes, we analysed the impact of the used bounding-boxes on the prediction performance of the networks.

On the DIPG dataset, we compared the object-detection performance with a generic bounding-box around the pons. This bounding-box was manually extracted from a template (Fonov et al., 2011) with an enlargement of approximately 50% on each side. Figure 5.3 summarises the experimental design chosen to evaluate the methods.

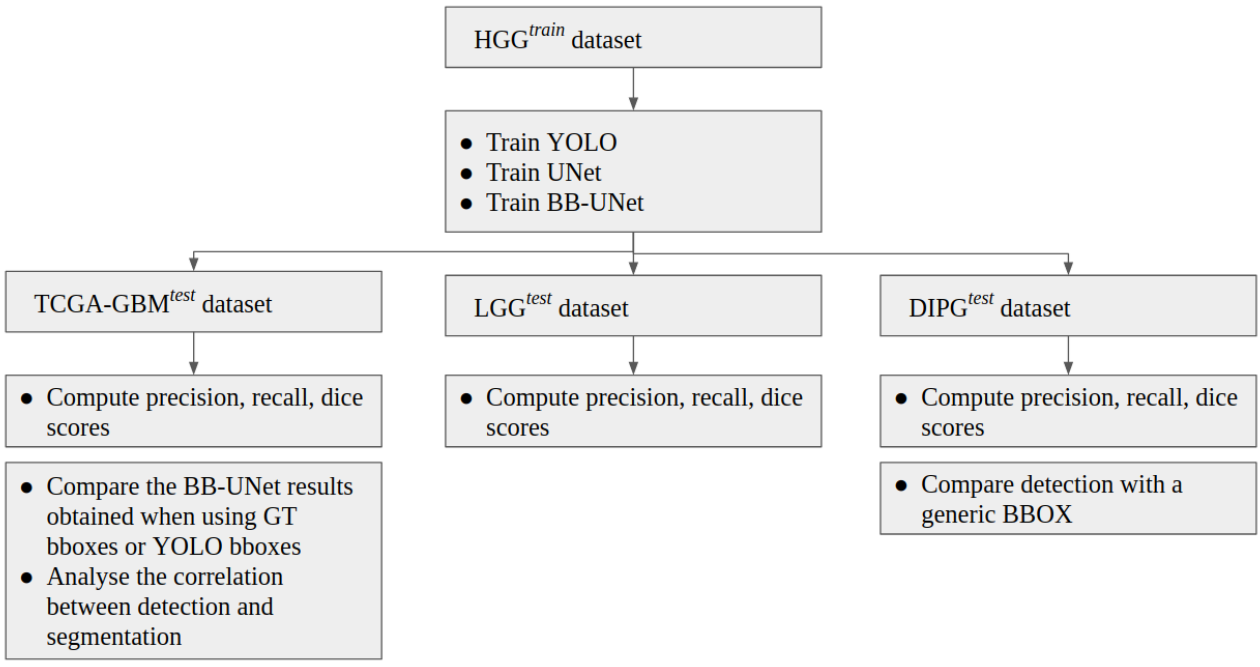


Figure 5.3: Experimental design diagram

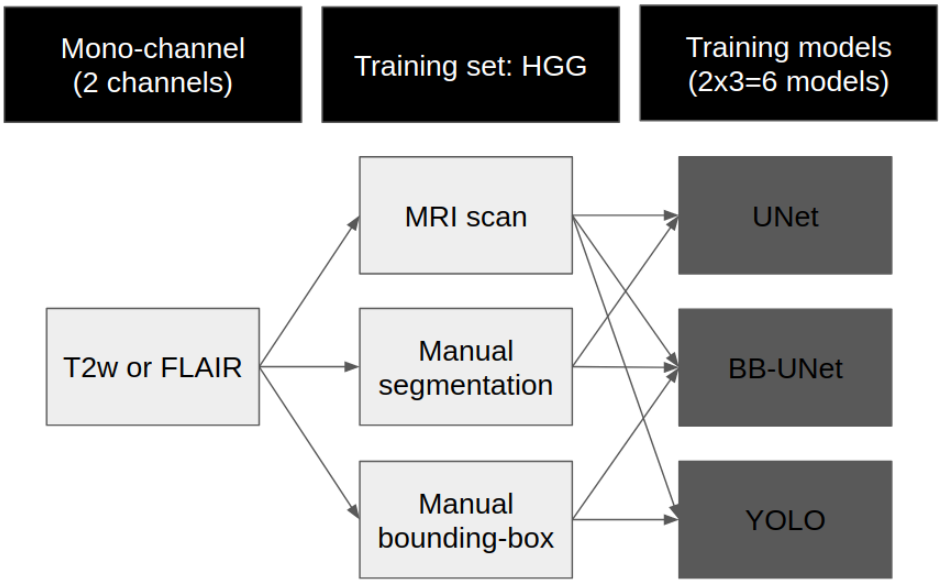


Figure 5.4: Details of the 6 different models that are trained for each modality. These 6 models will be used for inference either independantly, or serialised parallelized.

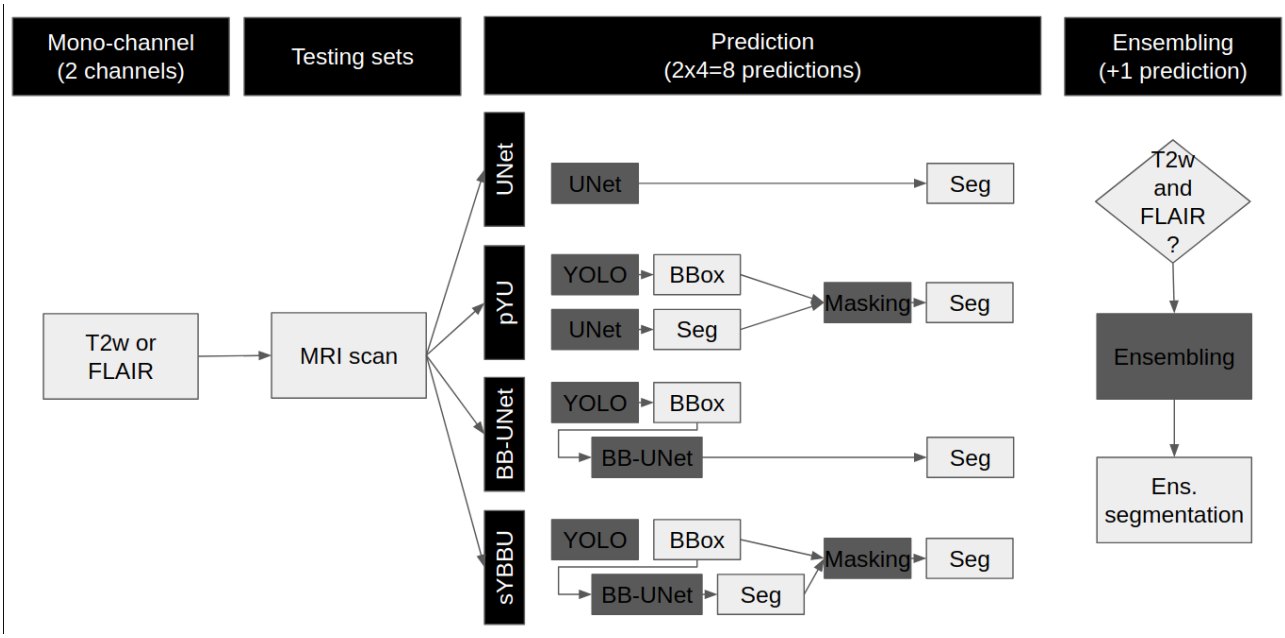


Figure 5.5: The trained models are used for inference either independently, or serialised parallelized. Finally results may be ensembled accross the modality.

5.5 Benchmark results

5.5.1 Object-detection results

Table 5.3 and 5.4 give the results of the detection phase on the TCGA-GBM and LGG datasets. Overall, both the FLAIR and the T2w obtain a very high recall and a relatively low precision score. The merging of both modalities helps further improve the recall and the stability of the predictions

(lowering the standard deviations) while lowering the precision. Low precision scores were expected in this phase since the predictions are piece-wise squares while tumour shapes are complex meshes. Therefore, the precision score depends heavily on the tumour shape and orientation. One must also note that a tumour generally occupies around 7% of the brain, in the studied dataset, which impacts the precision score. To choose the best model, it is important to remember that the main objective of this phase is to generate priors for a segmentation. It is therefore imperative to reliably detect the whole tumour (implying high recall), even if it comes with lower precision.

Table 5.3: Detection results on TCGA-GBM with 97 test patients. Results present the mean ± standard deviation

			With morphological transformation	
	Precision	Recall	Precision	Recall
FLAIR	.577(±.104)	.899(±.112)	.599(±.105)	.927(±.075)
T2w	.569(±.110)	.880(±.152)	.593(±.102)	.905(±.115)
ens.(FLAIR,T2w)	.511(±.103)	.945(±.069)	.527(±.105)	.956(±.059)

Table 5.4: Detection results on LGG with 76 test patients. Results present the mean ± standard deviation.

			With morphological transformation	
	Precision	Recall	Precision	Recall
FLAIR	.577(±.125)	.849(±.203)	.611(±.124)	.873(±.183)
T2w	.581(±.122)	.856(±.158)	.610(±.135)	.883(±.136)
ens.(FLAIR,T2w)	.503(±.121)	.926(±.121)	.529(±.123)	.940(±.094)

The detection framework achieved better performance in the TCGA-GBM dataset than in LGG. This was expected since our model was solely trained to detect high-grade-gliomas. Even if the performance was degraded for the LGG dataset, this decrease is moderate, especially when comparing the results of the ensembled model. This shows that the object-detection model is able to detect different types of tumours that occur in the same tissues of the brain.

Morphological opening and closing showed a minimal effect. However, as these effects were always positive on both precision and recall, we kept them in our detection process.

5.5.1.1 Segmentation results

YOLO bounding-boxes, obtained in the previous phase, were used during the segmentation. Each segmentation model uses bounding-boxes obtained from the same input image and modality. The bounding-boxes used for the segmentation are all post-processed by the morphological transformations.

Segmentation Results on TCGA-GBM Table 5.5 describes the results obtained for the segmentation of TCGA-GBM. A voxel is considered tumoural if its confidence score $\hat{y}_v$ is above 0.5. As expected, precision scores were considerably higher than during the detection phase, however, this came with a decrease in recall.

Table 5.5: Segmentation results on TCGA-GBM with 97 test patients. Unlike (BB-)UNet approaches, Deepmedic network (Kamnitsas et al., 2017) is trained on 3D volumes from the HGG^{train} dataset. Results present the mean ± standard deviation.

Architecture		Without masking				With masking		
		Precision	Recall	Dice		Precision	Recall	Dice
UNet	FLAIR	$.746 \pm .295$	$.825 \pm .100$	$.741 \pm .215$	pYU*	$.902 \pm .128$	$.813 \pm .103$	$.845 \pm .089$
	T2w	$.696 \pm .275$	$.823 \pm .163$	$.699 \pm .230$		$.858 \pm .140$	$.812 \pm .168$	$.813 \pm .134$
	ens.(FLAIR,T2w)	$.784 \pm .260$	$.843 \pm .103$	$.781 \pm .192$		$.914 \pm .115$	$.830 \pm .107$	$.861 \pm .088$
BB-UNet*	FLAIR	$.906 \pm .120$	$.809 \pm .091$	$.847 \pm .087$	sYBBU*	$.921 \pm .102$	$.807 \pm .093$	$.854 \pm .079$
	T2w	$.887 \pm .090$	$.807 \pm .127$	$.838 \pm .096$		$.901 \pm .083$	$.806 \pm .129$	$.843 \pm .094$
	ens.(FLAIR,T2w)	$.909 \pm .114$	$.834 \pm .094$	$.863 \pm .087$		$.925 \pm .096$	$.835 \pm .096$	$.869 \pm .079$
DeepMedic	FLAIR	$.913 \pm .110$	$.774 \pm .175$	$.820 \pm .149$				

* Models using YOLO bounding-boxes.

The UNet models performed poorly on TCGA-GBM compared to the other models. Indeed the mean Dice index ranged from 0.70 to 0.78, which is below the other models' averages, with values greater than 0.84. This was mainly due to their low precision scores. A deeper look into the results showed that UNet segments healthy bright spots of the brain the same way it segments the bright spots indicating the presence of the tumour, especially œdema. On the other side, UNet gave comparable results in the recall metric (i.e. percentage of the tumour detected).

Moreover, we can see a clear improvement after UNet segmentations were masked with the predicted bounding-boxes, i.e the pYU model. Mean precision scores were increased by nearly 15% in all configurations, while the standard deviations were reduced by nearly half. These results suggest that most of the false positives are outside of the bounding-boxes, which is in line with the assumption stated by the equation 5.1. These improvements came with a slight decrease in the recall, of around 1%. We consider this decrease is minor compared to the benefits of masking with bounding-boxes in precision.

Furthermore, bounding-boxes also have a positive effect when they are used as inputs in the BB-UNet models. There is an increase in mean precision and a decrease in standard deviations of precision and Dice index. After masking, the results of UNet and BB-UNet are very similar, with BB-UNet coming slightly ahead, with an improvement in precision between 1% and 3%, and a recall that remained similar among all the models. Figure 5.6 exhibits clearly that models using the bounding-boxes perform better, especially sYBBU models. To compare the sYBBU model as the best approach using bounding-box with UNet that ignores them, we computed the AUC of the mean precision-recall graph. We obtained 0.91, 0.90 and 0.93 on the FLAIR, T2w and ens.(FLAIR, T2w) respectively, on the sYBBU model. Meanwhile, on the UNet model, we obtained 0.80, 0.83 and 0.90 using the FLAIR, T2w and ens.(FLAIR, T2w) respectively.

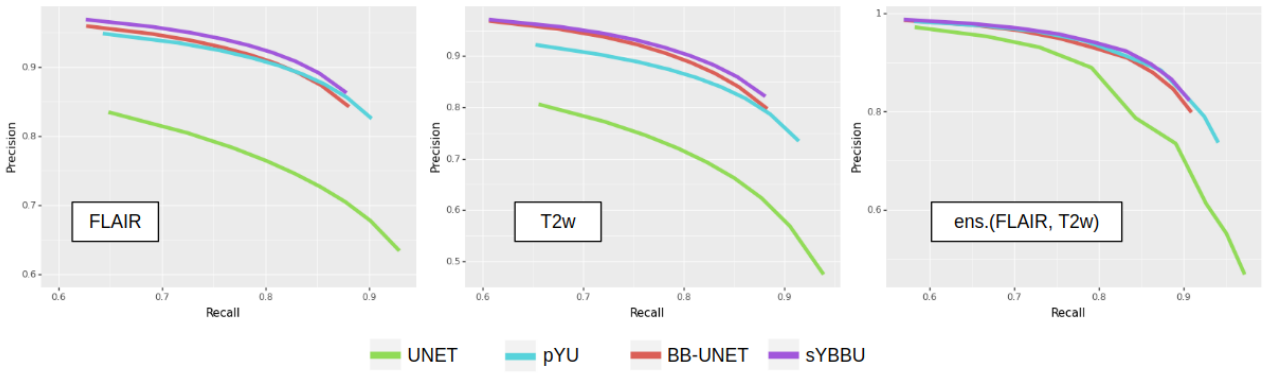


Figure 5.6: Mean precision-recall graphs of the different proposed segmentations on the TCGA-GBM dataset. To focus on the most interesting part of the plot, we only plotted the precision-recall scores for thresholds between 0.1 and 0.9. From the left to the right, Using the FLAIR, using the T2w, and ens.(FLAIR,T2w).

Table 5.6: Correlation study. Correlation values between detection precisions and final segmentation results obtained on the ensembled bounding-boxes.

	Precision	Recall	Dice
UNet	0.400	-0.083	0.334
pYU	0.665	-0.073	0.422
BB-UNet*	0.615	-0.037	0.363
sYBBU	0.682	-0.033	0.357

Table 5.7: Correlation study. Correlation values between detection recalls and final segmentation results obtained on the ensembled models.

	Precision	Recall	Dice
UNet	0.167	0.665	0.319
pYU	0.000	0.762	0.615
BB-UNet*	0.189	0.791	0.684
sYBBU	0.081	0.804	0.716

The precisions and recalls in segmentation using the bounding-boxes are strongly correlated with the respective precisions and recalls of the bounding-boxes detection, by a correlation score between 0.6 and 0.8 (see Tables 5.6 and 5.7). Of note, UNet results are also positively correlated with the object detection results (correlation between the precisions is 0.40 and correlation between the recalls is 0.66), though not as strongly as in the other models. This suggests that part of the performance is related to the images themselves, and some tumours are especially hard or easy to detect or segment for any model due to image quality or tumour visual characteristics. However, the overall performance is strongly dependent on YOLO's ability to detect the whole tumour. This is shown on the Dice metric, which indicates a strong correlation between bounding-box recall and the Dice of models using the bounding-boxes, ranging from 0.61 to 0.72. This reinforces the strategy consisting in promoting recall over precision during the detection phase in order to obtain overall high performance.

Overall, FLAIR-based models perform better than T2w-based models. It appears that the FLAIR may reflect the diffuse characteristics of the tumour better, while in the T2w images, the intensity

distribution of voxels inside the tumour is not as distinguishable from other bright regions of the brain. However, the ensembled models always perform better, across all configurations, and have equal or lower standard deviations. When computing the optimal weights to merge the models, we found $\omega = 0.77$ for pYU and $\omega = 0.50$ for sYBBU. However, the gain of an optimized weighted average, as opposed to a basic average, was below 1% (results not shown). The weighted average improves the log-likelihood, but with little impact on the accuracy after binarisation.

Table 5.8 shows the differences in metrics when ground-truth bounding-boxes were used for the FLAIR in BB-UNet. BB-UNet with ground-truth bounding-boxes was unable to detect, on average, 10% of the tumour. When YOLO bounding-boxes were used, a 6% decrease in recall was observed. This exhibits that two-thirds of the missed voxels are inherently related to BB-UNet and not to errors in YOLO bounding-boxes. Given these results, we can say that YOLO bounding-boxes are not the prevailing source of errors and they are sufficient to be integrated into our detection-segmentation approach.

Table 5.8: Comparison of segmentation performances when using the real bounding-boxes and predicted bounding-boxes for the FLAIR without post-processing. Results present the mean ± standard deviation.

	Precision	Recall	Dice
YOLO Bounding-Boxes	.906 ± .120	.809 ± .091	.847 ± .087
Real Bounding-Boxes	.932 ± .072	.875 ± .073	.899 ± .043

Concerning Deepmedic architecture, results on the FLAIR modality are slightly below results obtained with the proposed approaches. The Deepmedic model seems to prioritise high precision over recall. However, Deepmedic trained with the T2w failed to give any meaningful result, with an average Dice index of 0.08, which makes the T2w unusable in an ensembled model.

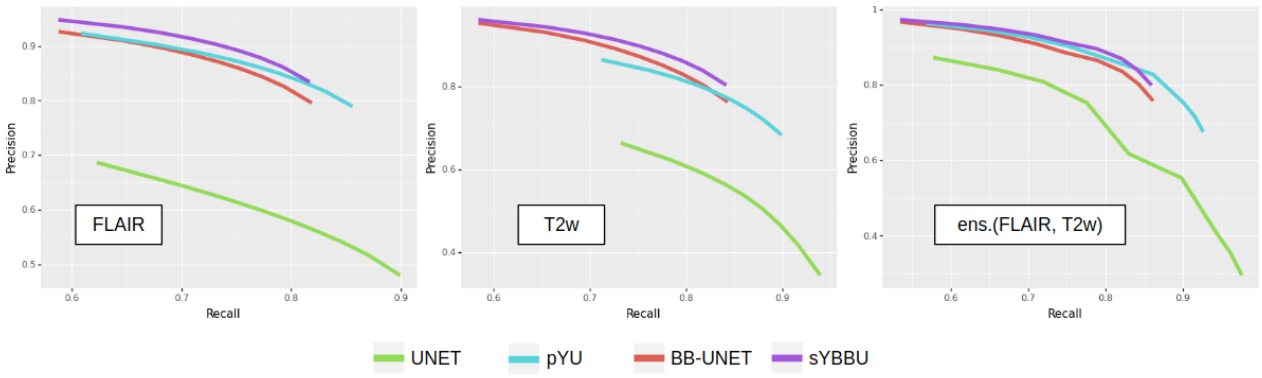


Figure 5.7: Mean precision-recall graphs of the different proposed segmentations on the LGG dataset. To focus on the most interesting region, we only plotted the precision-recall scores for thresholds between 0.1 and 0.9. From the left to the right, using the FLAIR, using the T2w, and ens.(FLAIR, T2w)

Segmentation Results on LGG Table 5.9 shows the segmentation results obtained on the LGG dataset. Overall, the proposed models exhibit comparable results with those obtained on the TCGA-GBM dataset but show an average drop in the Dice metric of 0.05. This reduction was expected since the networks were solely trained on High Grade Gliomas and were not readapted for the Low Grade

Glioma cases. Unlike the proposed models using bounding-boxes, UNet showed poor performances on the LGG dataset. However, the pYU model shows an improvement in the overall results by increasing precision by 30% at the cost of a mean decrease of 10% of the recall. We obtained an AUC score of 0.80, 0.87 and 0.89 for the FLAIR, T2w and ens.(FLAIR, T2w) respectively when using the sYBBU. Comparatively, we obtained an AUC score of 0.71, 0.73, 0.80 for the FLAIR, T2w and ens.(FLAIR, T2w) respectively when using UNet only. The precision-recall curves are given in Figure 5.7.

Table 5.9: Segmentation results on LGG with 76 test patients. Results present the mean ± standard deviation

Architecture		Without masking				With masking		
		Precision	Recall	Dice		Precision	Recall	Dice
UNet	FLAIR	$.541 \pm .360$	$.845 \pm .139$	$.577 \pm .296$	pYU*	$.863 \pm .155$	$.768 \pm .172$	$.792 \pm .142$
	T2w	$.467 \pm .305$	$.897 \pm .172$	$.523 \pm .295$		$.785 \pm .195$	$.831 \pm .195$	$.766 \pm .174$
	ens.(FLAIR,T2w)	$.345 \pm .275$	$.949 \pm .062$	$.444 \pm .303$		$.871 \pm .157$	$.800 \pm .160$	$.814 \pm .135$
BB-UNet*	FLAIR	$.844 \pm .165$	$.773 \pm .199$	$.790 \pm .169$	sYBBU*	$.878 \pm .144$	$.772 \pm .201$	$.804 \pm .165$
	T2w	$.828 \pm .144$	$.797 \pm .151$	$.801 \pm .150$		$.861 \pm .132$	$.796 \pm .181$	$.815 \pm .148$
	ens.(FLAIR,T2w)	$.835 \pm .161$	$.822 \pm 0182$	$.815 \pm .153$		$.871 \pm .141$	$.820 \pm .183$	$.831 \pm .152$
DeepMedic	FLAIR	$.904 \pm .146$	$.695 \pm .254$	$.743 \pm .215$				

* Models using YOLO bounding-boxes.

The proposed procedures outperformed the Deepmedic network. On average, the Dice metric was between 6% and 9% lower for Deepmedic compared to our models. This exhibits the robustness of our strategy. Similar to the TCGA-GBM dataset, the Deepmedic model seems to prioritise high precision (the highest of all the models) over recall.

5.5.2 Segmentation results on DIPG

Table 5.10: Detection results on 71 test cases from the DIPG set.

	Precision	Recall
FLAIR	0.529 ± 0.265	0.580 ± 0.313
T2w	0.436 ± 0.299	0.342 ± 0.341
ens.(FLAIR, T2w)	0.395 ± 0.229	0.667 ± 0.305

From the 30 DIPG patients, 71 sessions were available obtained at different follow-up visits. Table 5.11 shows the inference detection results obtained on 62 out of the 71 DIPG sessions. The detection step failed to identify the tumour region (tiny bounding-box with recall < 0.015) in 9 sessions (13% of the sample). An example of a failed detection is presented in Figure 5.8. Table 5.10 shows detection results obtained on all 71 test sessions. Overall, the FLAIR exhibited significantly better results than the T2w, especially for the recall. Bounding-boxes obtained from the FLAIR show robust results. On average, 66% of the tumour is captured, with a mean precision equal to 61%. While the T2w alone failed to give significant results, ensembling T2w bounding-boxes with FLAIR bounding-boxes improves the recall by 7% while lowering its precision by 17% on average. Both FLAIR and ens.(FLAIR, T2w) bounding-boxes performed better than the generic bounding-box around the pons, which gives a 63% recall with a 20% precision (results not shown). This shows that, even if the location of the tumour is known beforehand, the problem remains non-trivial because of the infiltrating nature of the tumour and its tendency to deform the surrounding tissue or structures (cerebellum, spinal

cord, thalamus). Careful inspection of the 9 cases with detection step failure indicates that the failure is mostly related to the tumours not being visible on the FLAIR and T2w MRI scans. Due to the low performance of object detection on T2w, we only used FLAIR and ens.(FLAIR, T2w) detection masks in the segmentation phase.

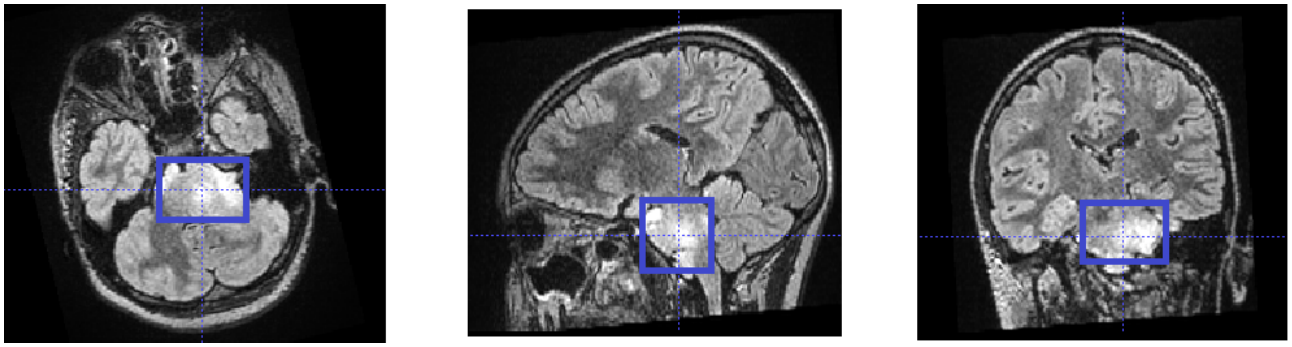


Figure 5.8: FLAIR image of a DIPG patient from the PREBIOMEDE cohort. Detection Failed on the image.

Table 5.11: Detection results on 62 sessions from the DIPG dataset excluding the 9 sessions where detection step failed to detect anything. Results present the mean ± standard deviation

	With morphological transformation	
	Precision	Recall
FLAIR	$.606 \pm .185$	$.664 \pm .237$
T2w	$.471 \pm .285$	$.363 \pm .334$
ens.(FLAIR,T2w)	$.447 \pm .189$	$.734 \pm .203$

Table 5.12: Segmentation results on 62 test sessions from the DIPG set, using ens.(FLAIR, T2w) detection masks.

		Precision	Recall	Dice
pYU	FLAIR	0.73 ± 0.20	0.59 ± 0.24	0.62 ± 0.21
	T2w	0.66 ± 0.19	0.63 ± 0.24	0.61 ± 0.19
	ens.(FLAIR, T2w)	0.67 ± 0.20	0.63 ± 0.24	0.61 ± 0.20
sYBBU	FLAIR	$.725 \pm .198$	$.609 \pm .205$	$.627 \pm .204$
	T2w	$.677 \pm .192$	$.596 \pm .203$	$.599 \pm .194$
	ens.(FLAIR, T2w)	$.688 \pm .192$	$.622 \pm .203$	$.618 \pm .197$

Table 5.13 shows pYU and sYBBU segmentation results using FLAIR detection masks. Table 5.12 shows segmentation results with the combined ens.(FLAIR,T2w) detection masks. Overall, across the configuration reported in Table 5.13, the mean Dice index for segmentation results is 61% (with 95% CI 0.56 to 0.66), which is satisfying considering the difficulty of the problem. An example of the segmentations obtained is presented in Figure 5.9.

Table 5.13: Segmentation results on 62 sessions from DIPG, using FLAIR detection masks. Results present the mean ± standard deviation

Architecture		Without masking				With masking		
		Precision	Recall	Dice		Precision	Recall	Dice
UNet	FLAIR	$.322 \pm .235$	$.644 \pm .225$	$.370 \pm .210$	pYU*	$.753 \pm .206$	$.571 \pm .243$	$.611 \pm .215$
	T2w	$.051 \pm .048$	$.749 \pm .232$	$.091 \pm .077$		$.680 \pm .192$	$.616 \pm .244$	$.606 \pm .202$
	ens.(FLAIR,T2w)	$.164 \pm .172$	$.703 \pm .234$	$.223 \pm .173$		$.695 \pm .205$	$.622 \pm .245$	$.609 \pm .212$
BB-UNet*	FLAIR	$.517 \pm .220$	$.699 \pm .236$	$.555 \pm .195$	sYBBU*	$.733 \pm .194$	$.598 \pm .244$	$.622 \pm .206$
	T2w	$.423 \pm .193$	$.701 \pm .231$	$.496 \pm .190$		$.697 \pm .195$	$.576 \pm .233$	$.596 \pm .200$
	ens.(FLAIR,T2w)	$.437 \pm .201$	$.719 \pm .235$	$.509 \pm .192$		$.711 \pm .195$	$.603 \pm .243$	$.616 \pm .202$
DeepMedic	FLAIR	$.624 \pm .225$	$0.614 \pm .259$	$.558 \pm .240$				

* Models using YOLO bounding-boxes.

Since the detection and segmentation phases can be done independently, we computed the performance of segmentation on the T2w, using FLAIR and ens.(FLAIR, T2w) detection masks. Indeed T2w detection, and segmentation using T2w did not fail. However, its results were still below FLAIR ones. On the T2w, pYU performance exhibits a dependence on the detection performance. Indeed, the pYU segmentation model failed to discriminate between tumoural voxels and healthy tissue ones, thus the segmentation results follow the detection performance. This is not the case for sYBBU, using ens.(FLAIR, T2w) detection masks, which have lower precision scores, and did not impact the segmentation model as much as the pYU. Looking at the Dice measurements, FLAIR and ens.(FLAIR, T2w) have similar performances whichever the bounding-boxes and the model used. However, FLAIR tends to have a higher precision while ens.(FLAIR, T2w) has a better recall. The choice between the two approaches should be made in regard to the application.

On the FLAIR, the Deepmedic network was outperformed by the detection model, and therefore obviously outperformed by the segmentation models that use the FLAIR mask. Deepmedic also failed to detect any tumour region in the same 9 cases excluded earlier.

5.6 Discussion

Our study proposes two detection-segmentation combination strategies that allowed us to obtain better results than the tested state-of-the-art networks (UNet and Deepmedic) on both BraTS'19, an openly available HGG and LGG dataset, and DIPG, a cohort of a rare paediatric tumour. Our strategies were able to segment the DIPG lesion while only training the models on the HGG cohort and without re-adapting the networks to the new tumour type. It was necessary to use this domain adaptation since we did not have access to enough annotated DIPG data nor a complete dataset to fine-tune each of the networks used.

Throughout this work, the FLAIR modality consistently appeared as the most important modality for any segmentation model, aiming at delineating globally the tumour lesion without distinguishing between its multiple compartments. It is therefore not surprising that our detection-segmentation algorithms prefer to rely on the FLAIR sequence. Moreover, the FLAIR modality has also been found as the most relevant for oncologists and features extracted from the FLAIR scans have shown the best results for survival analysis and tumour characterization for a range of tumours(Kickingereder et al., 2016). Specifically in DIPG, Castel *et al.* (Castel et al., 2015) identified differences in FLAIR index according to the type of histone mutated. Our segmentation, which is based on FLAIR imaging and

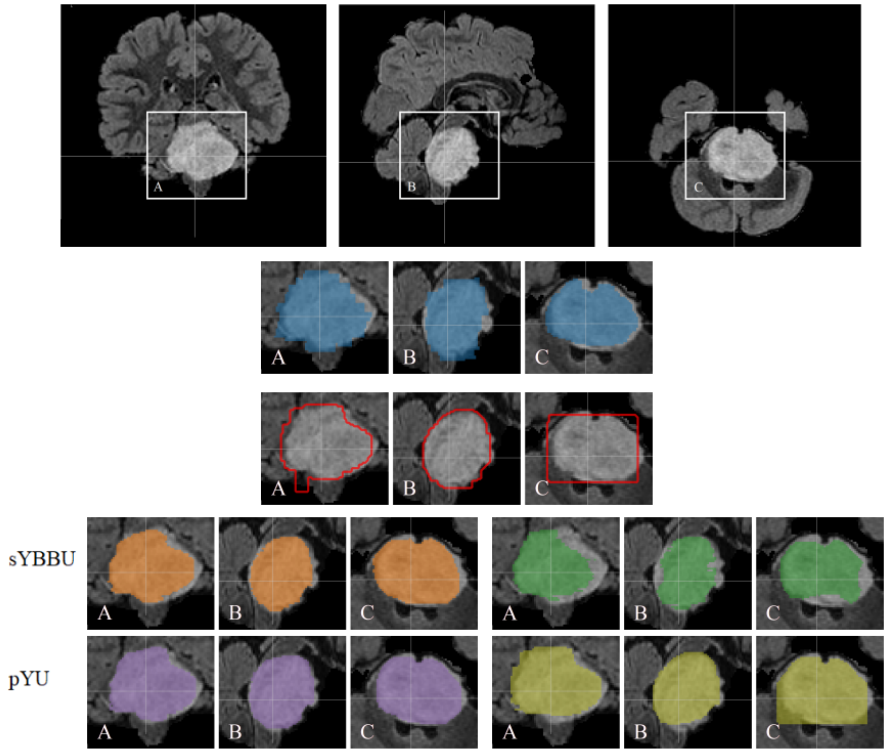


Figure 5.9: Segmentation results obtained on one case of DIPG are superimposed on the FLAIR background. The top row displays the complete patient images. Ground truth mask in blue. Yolo detection contours in red. sYBBU Segmentation with FLAIR in orange. sYBBU Segmentation with T2w in green. pYU Segmentation with FLAIR in purple. pYU Segmentation with T2w in yellow. The tumours are presented as a hyper-signal.

produces a FLAIR-mostly derived delineation, produces regions of interest that appear to be relevant. In addition to that it appears that, even if the T2w did not perform as well as expected for the DIPG dataset, its presence always helped the proposed segmentation models.

Our proposals consist of procedures implicating multiple different and distinct models. Having different models, trained separately, has several advantages. The models had very different architectures, and therefore, could have different weaknesses and strengths, which can be complementary. In the DIPG case, even when the T2w detection failed, we were able to use the trained T2w BB-UNet model efficiently using the alternate FLAIR bounding-boxes. This possibility allowed us to circumvent the differences between glioblastoma and DIPG.

Our proposed approaches consist in combining multiple models, each model is relatively small. The inference time for each 3D example is around 5 seconds on an Nvidia Titan X, including 2.5 seconds for detection and 2.5 seconds for segmentation. Meanwhile, the training phase took roughly 4 hours each. Comparatively, training Deepmedic took 3 days and had an inference time of 3 minutes per 3D example, on similar machine and software configurations.

Most recent segmentation efforts have focused on developing deeper and more complex models. While these solutions can be suitable for tumour lesions for which large curated and well-documented datasets, there is no indication that they can be easily adapted to small cohorts of rare tumours, such as DIPG, with missing data and heterogeneous quality. We found that Deepmedic, trained with four

modalities (FLAIR, T2w, T1w, T1wgd), performed exceptionally well for HGG, with an average Dice of 0.9, but fails on DIPG with an average Dice of 0.3. In our proposition, the segmentation model is not fully dependent on the object detection performance, given bounding-boxes can be obtained from other input images. This allows us to use the best bounding-boxes assessed during a quality check. Our results are in line with the work of Isensee et al. (2021), which found that recent efficient very complex and deep networks cannot necessarily be easily fine-tuned for rare oncological lesions segmentation problems with few training examples and, promoted the UNet architecture.

Our study presents several limitations. The ground-truth segmentations obtained on the DIPG are done on thick slices of $4mm^3$, which negatively bias the results obtained even if it does not question the magnitude of these results. Furthermore, we have only one set of rare tumour data, further studies should investigate the robustness of the method using other rare tumours. Additionally, throughout this study, the used networks are considered black boxes. An ablation study could be made to investigate the limitation of the networks. We also did not investigate what the networks learnt and how they make the inference. Finally, we only focused on binary segmentation using either the FLAIR or T2w, further studies should investigate multi-compartment segmentation possibly using other modalities.

5.7 Conclusion

This paper addresses the problem of rare tumour types, for which no database can be built to train a deep neural segmentation network. Our work shows that state-of-the-art segmentation methods perform poorly when applied on test cohorts on which they were not trained. We propose to combine different simple models of detection and segmentation to allow us, not only to improve UNet performance but also to obtain satisfying results on a cohort that contained differences compared to the training dataset regarding, among others, patient age, image quality and tumour type.

Although all the sets presented in the paper concern cerebral tumours, the differences between an adult brain (in the case of the HGGs and LGGs) and children's brains (in the case of DIPG) give rise to challenges during inference. We think that using a set of a wider range of brain tumour types in children might help solve this issue. Additionally, the paper does not explore alternatives to the object detection framework YOLO. Work should be done to compare it to other algorithms, especially the ones dedicated to medical imaging and not only natural images. Lastly, other detection-segmentation strategies, such as the weak supervision paradigm, can be explored and compared to the proposed approaches.

Overall, the bounding boxes give useful *a priori* knowledge for segmentation purposes. We have already established that these bounding boxes, when correctly detecting the tumour, can be used for radiomic analysis with similar performances to radiomics extracted using fine delineation of the tumour Chegraoui et al. (2021b). In this work, we were able to obtain satisfying segmentation for the DIPG. These segmentations and performance will allow us to perform further clinical work to characterise this rare pathology using radiomics.

* * *
* *
*

Chapter 6

Multiblock Integration Method

In the medical field, statistical models are often used to discover and identify variables of interest that influence the development or status of the studied pathology. These statistical models not only predict the outcome but also links a set of variables of interest to it. However, genes interact in complex patterns. Some genes regulate others, and some genes are co-expressed with each other. These interaction patterns are described in gene-gene interaction graphs and can be used in multiple ways in the biostatistical field. Including these graphs in the statistical models is one way to use them, highlighting not only the variables of interest but also the biological interaction of interest.

Furthermore, a new class of statistical models emerged as multiple measurements became available for the same patients from various sources. These statistical models differentiate between the different sources of data in separate blocks. The goal was expanded to understand the interactions between the different sources of data.

This work focuses on the addition of a graph penalty to a Regularized Generalized Canonical Correlation Analysis (RGCCA) model. The chapter starts with a mathematical definition of the RGCCA. Then we describe the variable selection process of the RGCCA. Later we present the RGCCA with a graph penalty.

Throughout the chapter, we define our models on a set of J blocks $X^{(1)}, X^{(2)}, \cdots, X^{(J)}$. Each block consists in data from the same sample of n patients and $p^{(j)}$ variables. The blocks are assumed to be centred and scaled.

Availability: Source code is freely available at `https://github.com/neurospin/netSGCCA`

Chapter Outline

Contents

6.1 Regularized Generalized Canonical Correlation Analysis . . . 71
6.2 Sparse Generalized Canonical Correlation Analysis . . . 72
6.3 netSGCCA . . . 72
6.3.1 GraphNet . . . 73
6.3.2 Optimisation . . . 74
6.3.3 The proximal operator . . . 75
6.4 Application . . . 76
6.4.1 Simulated Data . . . 76
6.4.2 TCGA-LGG analysis design and results . . . 80
6.4.3 Discussion . . . 88
6.4.4 Conclusion . . . 89

6.1 Regularized Generalized Canonical Correlation Analysis

The RGCCA was proposed by Tenenhaus and Tenenhaus (2011) and extended the classical Canonical Correlation Analysis (CCA) to a set of J blocks. The model aims at extracting shared information between the blocks using the following optimisation problem :

$$\begin{aligned} &\underset{\boldsymbol{w}^{(1)},\boldsymbol{w}^{(2)},\cdots,\boldsymbol{w}^{(J)}}{\operatorname{argmin}} \sum_{\substack{j,k=1 \\ j\neq k}}^{J} -c_{j,k}\, g\left(\frac{1}{n}\boldsymbol{w}^{(j)\top}\boldsymbol{X}^{(j)\top}\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}\right) \\ &\text{s.t.} \boldsymbol{w}^{(j)} \in \Omega^{(j)}_{RGCCA} \quad j = 1,\cdots,J \end{aligned} \tag{6.1}$$

The coefficients $c_{j,k}$ are equal to 1 or 0 and define if the blocks k and j are connected. They can be seen as the coefficients of the adjacency matrix of an *a priori* design graph of interactions among the blocks.

The function g can be any real continuous differentiable function. Authors of the RGCCA propose the identity $g(x) = x$, the absolute value $g(x) = |x|$ or the square function $g(x) = x^2$. Only the identity penalises the negative correlation between the blocks.

$\Omega^{(j)}_{RGCCA}$ defines a set of constraints on each block weights. Here $\Omega^{(j)}_{RGCCA} = \{\boldsymbol{w}^{(j)} \in \mathbb{R}^{p^{(j)}}; \boldsymbol{w}^{(j)\top}\boldsymbol{M}_j\boldsymbol{w}^{(j)} = 1\}$, where $\boldsymbol{M}_j = \tau^{(j)} I_{p^{(j)}} + \frac{1-\tau^{(j)}}{n-1}\boldsymbol{X}^{(j)\top}\boldsymbol{X}^{(j)}$. $\tau^{(j)}$ is a shrinkage parameter between 0 and 1. $\tau^{(j)} = 1$ means the projectors ($\boldsymbol{w}^{(j)}$) must have a unit norm, while $\tau^{(j)} = 0$ means the projections ($\boldsymbol{X}^{(j)}\boldsymbol{w}^{(j)}$) must have a unit variance and the optimisation problem 6.1 aims at maximising the correlation between the blocks.

The weights $\boldsymbol{w}^{(j)}$ are estimated by maximising the sum of the correlations between pairs of latent components. Cycle block coordinate ascent, described in Algorithm 1, is used to solve the optimisation problem defined in 6.1. This algorithm updates each block weight in turn, keeping the others fixed (line 8 in Algorithm 1). Using the Lagrangian multipliers, the solution of the optimisation problem for block $\boldsymbol{X}^{(j)}$, if all other block parameters are considered fixed, is given by the Equation 6.2.

Algorithm 1 RGCCA optimisation algorithm

Require: $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \cdots, \boldsymbol{X}^{(J)}$, $\tau^{(1)}, \tau^{(2)}, \cdots, \tau^{(J)}$, design matrix C
Ensure: Optimal $\boldsymbol{w}^{(1)\star}, \cdots, \boldsymbol{w}^{(J)\star}$
1: **for** $j \in 1, \cdots, J$ **do**
2: $\quad \boldsymbol{w}^{(j)}_0$ is a random vector or the first singular vector of $\boldsymbol{X}^{(j)}$
3: $\quad$ Ensure $\boldsymbol{w}^{(j)}_0 \in \Omega^{(j)}_{RGCCA}$
4: **end for**
5: **repeat**
6: $\quad$ **for** $j \in 1, \cdots, J$ **do**
7: $\quad\quad \boldsymbol{z}^{t+1}_j = \sum_{k=1}^{j} c_{j,k} g'\left(\boldsymbol{w}^{(j)\top}_t \boldsymbol{X}^{(j)\top}\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}_{t+1}\right)\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}_{t+1} +$
$\quad\quad\quad \sum_{k=j+1}^{J} c_{j,k} g'\left(\boldsymbol{w}^{(j)\top}_t \boldsymbol{X}^{(j)\top}\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}_{t}\right)\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}_{t}$
8: $\quad\quad \boldsymbol{w}^{(j)}_{t+1} = \frac{\frac{1}{n}\boldsymbol{M}_j^{-1}\boldsymbol{X}^{(j)}\boldsymbol{z}_j}{\|\frac{1}{n}\boldsymbol{M}_j^{-1}\boldsymbol{X}^{(j)}\boldsymbol{z}^{t+1}_j\|}$
9: $\quad$ **end for**
10: **until** $\|\boldsymbol{w}^{(1)}_{t+1} - \boldsymbol{w}^{(1)}_t\| \leq \epsilon, \cdots, \|\boldsymbol{w}^{(J)}_{t+1} - \boldsymbol{w}^{(J)}_t\| \leq \epsilon$

$$\begin{aligned} \boldsymbol{w}^{(j)} &= \frac{\frac{1}{n}\boldsymbol{M}_j^{-1}\boldsymbol{X}^{(j)}\boldsymbol{z}_j}{\|\frac{1}{n}\boldsymbol{M}_j^{-1}\boldsymbol{X}^{(j)}\boldsymbol{z}_j\|} \\ \text{s.t.} \quad \boldsymbol{z}_j &= \sum_{k=1}^{J} c_{j,k} g'\left(\boldsymbol{w}^{(j)\top}\boldsymbol{X}^{(j)\top}\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}\right)\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)} \end{aligned} \tag{6.2}$$

6.2 Sparse Generalized Canonical Correlation Analysis

As defined in equation 6.1, the RGCCA does not allow for the identification of variables of interest among a large set of variables in each block. To tackle this issue, the Sparse Generalized Canonical Correlation Analysis (SGCCA) was introduced (Tenenhaus et al., 2014). The authors extended the RGCCA by adding a sparsity constraint on the model weights, using an ℓ_1 penalisation. The set of constraints for the SGCCA becomes $\Omega^{(j)}_{SGCCA} = \Omega^{(j)}_{RGCCA} \cap \{\boldsymbol{w}^{(j)} \in \mathbb{R}^{p^{(j)}}; \|\boldsymbol{w}^{(j)}\|_1 \leq s^{(j)}\}$, where $s^{(j)}$ is a user-chosen parameter. Additionally, the authors also fixed the shrinkage parameter $\tau^{(j)}$ in the RGCCA constraints to 1 for all the J blocks. Thus, the RGCCA constraints becomes $\Omega^{(j)}_{RGCCA} = \{\boldsymbol{w}^{(j)} \in \mathbb{R}^{p^{(j)}}; \|\boldsymbol{w}^{(j)}\|_2 = 1\}$.

Since $\|\boldsymbol{w}^{(j)}\|_2 \leq \|\boldsymbol{w}^{(j)}\|_1 \leq \sqrt{p^{(j)}}\|\boldsymbol{w}^{(j)}\|_2$ holds for all vectors, $\Omega^{(j)}_{SGCCA}$ is not empty if and only if $s^{(j)} \geq 1$. Additionally, the ℓ_1 is active if and only if $s^{(j)} \leq \sqrt{p^{(j)}}$.

Using the Lagrangian multipliers, the authors demonstrated that the optimal solution has the form $\boldsymbol{w}^{(j)} = \frac{S(\frac{1}{n}\boldsymbol{X}^{(j)\top}\boldsymbol{z}_j, \lambda^{(j)})}{\|S(\frac{1}{n}\boldsymbol{X}^{(j)\top}\boldsymbol{z}_j, \lambda^{(j)})\|_2}$, where S denotes the soft-thresholding operator and $\lambda^{(i)}$ is chosen such that $\|\boldsymbol{w}^{(j)}\|_1 \leq s^{(j)}$. The optimisation algorithm is similar to the one introduced in Algorithm 1, where the update of $\boldsymbol{w}^{(j)}$, in line 7, takes the new form.

6.3 netSGCCA

This section describes a modification to the SGCCA that injects prior graphical knowledge into the model. Given a graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$, where $\mathcal{E}$ is the set of a block $\boldsymbol{X}^{(g)}$ variables, we aim at imposing the graph structure on the block coefficients. The objective is to reflect the basic intuition that interacting variables (neighbouring variables in the graph) should contribute similarly to the model.

Following the work of Guigui et al. (2019), we propose to extend the SGCCA to netSGCCA by adding a GraphNet penalty on one block, denoted $\boldsymbol{X}^{(g)}$, relaxing the ℓ_2 equality constraint and using a gradient descent method to optimise the problem. The extension of our proposal to penalise more than one block is straightforward. Given the graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$, we introduce its Laplacian $\boldsymbol{L}_{\mathcal{G}}$ (or its normalised Laplacian) into the SGCCA optimisation problem stated, which then becomes:

$$\begin{aligned} &\underset{\boldsymbol{w}^{(1)},\boldsymbol{w}^{(2)},\cdots,\boldsymbol{w}^{(J)}}{\operatorname{argmin}} \sum_{\substack{j,k=1 \\ j\neq k}}^{J} -c_{j,k}\operatorname{cov}\left(\boldsymbol{X}^{(j)}\boldsymbol{w}^{(j)}, \boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}\right) + \frac{\gamma_{\mathcal{G}}}{\lambda_{\max}}\boldsymbol{w}^{(g)\top}\boldsymbol{L}_{\mathcal{G}}\boldsymbol{w}^{(g)} \\ &\text{s.t.} \quad \boldsymbol{w}^{(j)} \in \Omega^{(j)}_{netSGCCA} \quad j = 1, \cdots, J \end{aligned} \tag{6.3}$$

The binary symmetric design matrix $C = (c_{j,k})$ is the same as in the RGCCA optimisation problem. The constraints sets $\Omega^{(j)}_{netSGCCA} = \{\boldsymbol{w}^{(j)} \in \mathbb{R}^{p^{(j)}}; \|\boldsymbol{w}^{(j)\top}\|_2 \leq 1\} \cap \{\boldsymbol{w}^{(j)} \in \mathbb{R}^{p^{(j)}}; \|\boldsymbol{w}^{(j)}\|_1 \leq s^{(j)}\}$. As in the SGCCA optimisation problem, the $s^{(j)}$ are the sparsity parameters.

The Karush–Kuhn–Tucker conditions state that if the inequality constraint of an optimisation problem is active, the optimal solution necessarily satisfies the corresponding equality constraint. As explained by Witten et al. (2009), if the chosen sparsity constraints $s^{(j)}$ lead to an active ℓ_2 inequality constraint, the constraint sets $\Omega_{netSGCCA}$ and Ω_{SGCCA} are equivalent. We kept the relaxation of the ℓ_2 equality (i.e. we use the inequality in our formulas) as it leads to a convex optimisation problem with a convex set of constraints.

6.3.1 GraphNet

The Laplacian matrix associated with a graph $\mathcal{G} = (\mathcal{E}, \mathcal{V})$ is defined as $L_\mathcal{G} = D - A$, where D is the diagonal degrees matrix, and A is the graph adjacency matrix. Using these matrices and for any vector $\boldsymbol{w}$ of size p, the GraphNet penalty takes the form :

$$\boldsymbol{w}^\top L_\mathcal{G} \boldsymbol{w} = \sum_{\substack{j,k=1 \\ (k,j)\in\mathcal{V}}}^{p} (\boldsymbol{w}_j - \boldsymbol{w}_k)^2 \tag{6.4}$$

The Normalised Laplacian associated with the graph is defined as $L^n_\mathcal{G} = D^{-\frac{1}{2}} L_\mathcal{G} D^{-\frac{1}{2}}$. Then, with d_j the degree of the node j, the GraphNet penalty becomes :

$$\boldsymbol{w}^\top L^n_\mathcal{G} \boldsymbol{w} = \sum_{\substack{j,k=1 \\ (k,j)\in\mathcal{V}}}^{p} (\frac{\boldsymbol{w}_j}{\sqrt{d_j}} - \frac{\boldsymbol{w}_k}{\sqrt{d_k}})^2 \tag{6.5}$$

Looking at these formulations, we can see that GraphNet explicitly attempts to give similar weights to variables connected in the graph. Using the normalised Laplacian, one can expect variables with more neighbours to have higher weight values, which can be more stable. The higher weights for higher connected variables can be explained because, supposing two variables j, k with $d_j \geq d_k$, GraphNet will try to impose $\boldsymbol{w}_j = \frac{\sqrt{d_j}}{\sqrt{d_k}} \boldsymbol{w}_k$, with $\frac{\sqrt{d_j}}{\sqrt{d_k}} \geq 1$.

GraphNet uses a symmetric, semi-defined positive matrix. Therefore, the largest eigenvalue $\lambda_{\max}$ of this matrix defines the upper-bound of the penalty (and its gradient), i.e., $\boldsymbol{w}^\top \boldsymbol{L}_\mathcal{G} \boldsymbol{w} \leq \lambda_{\max} |\boldsymbol{w}|_2^2$. The hyper-parameter $\gamma_\mathcal{G}$, introduced in the equation 6.3 defines the importance of GraphNet in the objective function. By dividing the GraphNet penalty by its largest eigenvalue, $\gamma_\mathcal{G}$ is comparable between different Laplacians, which will make clearer the benchmarks of graphs we will present thereafter.

The GraphNet penalty has a grouping effect, which means that the weights of linked variables are brought closer. To prove the grouping effect, we can rewrite the equation 6.3, using the lagrangian multipliers $\lambda_1^{(j)}$ and $\lambda_2^{(j)}$ respectively associated with the ℓ_1 and ℓ_2 constraints for each block:

$$\begin{aligned} &\underset{\boldsymbol{w}^{(1)},\boldsymbol{w}^{(2)},\cdots,\boldsymbol{w}^{(J)}}{\operatorname{argmin}} \; h(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(1)}, \cdots, \boldsymbol{w}^{(J)}) \\ h(\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \cdots, \boldsymbol{w}^{(J)}) &= \sum_{\substack{j,k=1 \\ j\neq k}}^{J} -c_{j,k} \operatorname{cov}\left(\boldsymbol{X}^{(j)}\boldsymbol{w}^{(j)}, \boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}\right) + \frac{\gamma_\mathcal{G}}{\lambda_{\max}} \boldsymbol{w}^{(g)\top} \boldsymbol{L}_\mathcal{G} \boldsymbol{w}^{(g)} \\ &\quad + \sum_{j=1}^{J} \lambda_1^{(j)} \|\boldsymbol{w}^{(j)}\|_1 + \sum_{j=1}^{J} \lambda_2^{(j)} \|\boldsymbol{w}^{(j)}\|_2^2 \end{aligned} \tag{6.6}$$

For two variables u and v, only connected to each other in the graph and positively correlated, we can show (see Appendix B) that:

$$|\boldsymbol{w}_u^{(g)} - \boldsymbol{w}_v^{(g)}| \leq \frac{Q\sqrt{2(1 - corr(X_u^{(g)}, X_v^{(g)}))}}{2\left(\frac{\gamma_{\mathcal{G}}}{\lambda_{\max}} + \lambda_2^{(g)}\right)} \tag{6.7}$$

Q is a constant only dependent on the correlation matrix of the other blocks. The proof can be found in the annex, and it follows the demonstration established by Li and Li (2008) for the grouping effect of the GraphNet in the context of the regression.

6.3.2 Optimisation

As proposed by Löfstedt et al. (2016), we used the Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) (Beck and Teboulle, 2009) to find the optimal solution to our optimisation problem. The FISTA is built on the Iterative Shrinkage-Thresholding Algorithm (ISTA), which is designed to solve optimisation problems of the form:

$$\underset{\boldsymbol{x}}{\operatorname{argmin}} \{F(\boldsymbol{x}) = f(\boldsymbol{x}) + g(\boldsymbol{x})\} \tag{6.8}$$

where the functions f and g verify: f is a smooth real convex function continuously differentiable with Lipschitz continuous gradient (i.e. $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\| \leq L(f)\|\boldsymbol{x} - \boldsymbol{y}\|$ for every $\boldsymbol{x}, \boldsymbol{y}$), and g is a real continuous convex function and possibly non-smooth.

Instead of the classic gradient descent, the ISTA uses the equivalent proximal formulation. Starting from an initial point $\boldsymbol{x}_0$, the ISTA generates a sequence $\boldsymbol{x}_k$ verifying:

$$\boldsymbol{x}_{k+1} = \text{prox}_g(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)) \tag{6.9}$$

where the proximal operator is defined as $\text{prox}_g(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{y}} \{g(\boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2\}$ and α_k is a positive real number.

To accelerate the convergence rate of the ISTA, the FISTA uses Nesterov's accelerated gradient descent. The authors have shown that the FISTA converges at a rate $\mathcal{O}(\frac{1}{k^2})$, and it is described as follows, with $\theta_0 = 0$:

$$\begin{aligned}
\theta_{k+1} &= \frac{1 + \sqrt{1 + 4\theta_k^2}}{2} \\
\beta_{k+1} &= \frac{1 - \theta_k}{\theta_{k+1}} \\
\boldsymbol{x}'_{k+1} &= \text{prox}_g(\boldsymbol{x}_k - \alpha_k \nabla f(\boldsymbol{x}_k)) \\
\boldsymbol{x}_{k+1} &= (1 - \beta_{k+1})\boldsymbol{x}'_{k+1} + \beta_{k+1}\boldsymbol{x}'_k
\end{aligned} \tag{6.10}$$

We solve the optimisation problem defined in 6.6 sequentially, i.e. we find the optimal solution for each block while fixing the parameters related to the other blocks. For a block i, fixing the other blocks parameters results in the following optimisation problem:

$$
\begin{aligned}
F(\boldsymbol{w}^{(i)}) &= f(\boldsymbol{w}^{(i)}) + g(\boldsymbol{w}^{(i)}) \\
f(\boldsymbol{w}^{(i)}) &= \sum_{\substack{k=1 \\ k\neq i}}^{J} -c_{i,k}\,\mathrm{cov}\left(\boldsymbol{X}^{(i)}\boldsymbol{w}^{(i)}, \boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}\right) + \frac{\gamma_{\mathcal{G}}}{\lambda_{\max}}\boldsymbol{w}^{(i)\top}\boldsymbol{L}_{\mathcal{G}}\boldsymbol{w}^{(i)} \\
g(\boldsymbol{w}^{(i)}) &= \lambda_1^{(i)}\|\boldsymbol{w}^{(i)}\|_1 + \lambda_2^{(i)}\|\boldsymbol{w}^{(i)}\|_2^2
\end{aligned}
\tag{6.11}
$$

The function f and g satisfy the FISTA conditions and the optimisation algorithm is described in 2. Since f has a Lipschitz continuous gradient, the gradient descent step α_k is fixed at $\alpha_k = \frac{1}{L}$ with L the Lipschitz constant of the gradient.

Algorithm 2 netSGCCA optimisation algorithm

Require: $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \cdots, \boldsymbol{X}^{(J)}$, $\boldsymbol{L}_{\mathcal{G}}$, $\gamma_{\mathcal{G}}$, design matrix C
Ensure: Optimal $\boldsymbol{w}^{(1)\star}, \cdots, \boldsymbol{w}^{(J)\star}$
for $j \in 1, \cdots, J$ **do**
  $\boldsymbol{w}_0^{(j)}$ is a random vector or the first singular vector of $\boldsymbol{X}^{(j)}$
  Ensure $\boldsymbol{w}_0^{(j)} \in \Omega_{netSGCCA}^{(j)}$
end for
repeat
  for $j \in 1, \cdots, J$ **do**
    repeat
      $\theta_{k+1} = \frac{1+\sqrt{1+4\theta_k^2}}{2}$
      $\beta_{k+1} = \frac{1-\theta_k}{\theta_{k+1}}$
      $\boldsymbol{w}_{k+1}^{(j)'} = \mathrm{prox}_g(\boldsymbol{w}_k^{(j)} - \alpha_k \nabla f(\boldsymbol{w}_k^{(j)}))$
      $\boldsymbol{w}_{k+1}^{(j)} = (1-\beta_{k+1})\boldsymbol{w}_{k+1}^{(j)'} + \beta_{k+1}\boldsymbol{w}_k^{(j)'}$
    until $\|\boldsymbol{w}_{k+1}^{(j)} - \boldsymbol{w}_k^{(j)}\| \leq \epsilon$
  end for
until $\|f^k(\boldsymbol{w}^{(1)}) - f^{k+1}(\boldsymbol{w}^{(1)})\| \leq \epsilon, \cdots, \|f^k(\boldsymbol{w}^{(J)}) - f^{k+1}(\boldsymbol{w}^{(J)})\| \leq \epsilon$

6.3.3 The proximal operator

In the netSGCCA, the proximal operator can be written as $proj_g(\boldsymbol{x}) = \mathrm{argmin}_{\boldsymbol{y}\in\Omega_{netSGCCA}}\{\frac{1}{2}\|\boldsymbol{x}-\boldsymbol{y}\|_2^2\}$. This defines a projection into a convex set problem, which can be solved using the Dykstra projection algorithm (Bauschke and Borwein, 1994).

The Dykstra algorithm is designed for the projection on the intersection of two convex sets. Let $\mathcal{S}$ and $\mathcal{R}$ be two convex sets, with $proj_{\mathcal{S}}$ and $proj_{\mathcal{R}}$ their respective projection operators, the Dykstra algorithm generates a sequence $\boldsymbol{x}_k$ converging into the projection of $\boldsymbol{x}$ onto $\mathcal{S}\cap\mathcal{R}$. Starting from $\boldsymbol{x}_0 = \boldsymbol{x}$ and with $\boldsymbol{p}_0 = \boldsymbol{q}_0 = 0$, the algorithm is described as follows:

$$
\begin{aligned}
\boldsymbol{y}_k &= proj_{\mathcal{S}}(\boldsymbol{x}_k + \boldsymbol{p}_k) \\
\boldsymbol{p}_{k+1} &= \boldsymbol{x}_k + \boldsymbol{p}_k - \boldsymbol{y}_k \\
\boldsymbol{x}_{k+1} &= proj_{\mathcal{R}}(\boldsymbol{y}_k + \boldsymbol{q}_k) \\
\boldsymbol{q}_{k+1} &= \boldsymbol{y}_k + \boldsymbol{q}_k - \boldsymbol{x}_{k+1}
\end{aligned}
\tag{6.12}
$$

In the netSGCCA, $\mathcal{S} = \{\boldsymbol{w}^{(j)} \in \mathbb{R}^{p_j}; \|\boldsymbol{w}^{(j)}\|_1 \leq s^{(j)}\}$ and $\mathcal{R} = \{\boldsymbol{w}^{(j)} \in \mathbb{R}^{p_j}; \|\boldsymbol{w}^{(j)}\|_2 \leq 1\}$.

The $proj_{\mathcal{S}}(\boldsymbol{x})$ is the soft-thresholding operator defined as:

$$(proj_{\mathcal{S}}(\boldsymbol{x}))_i = sign(\boldsymbol{x}_i)max(|\boldsymbol{x}_i| - \lambda, 0) \tag{6.13}$$

Where λ is a positive scalar chosen to satisfy the constrain set $\mathcal{S}$.

The $proj_{\mathcal{R}}(\boldsymbol{x})$ is defined as:

$$proj_{\mathcal{R}}(\boldsymbol{x}) = \begin{cases} \frac{1}{\|\boldsymbol{x}\|_2}\boldsymbol{x} & \|\boldsymbol{x}\|_2 \geq 1 \\ \boldsymbol{x} & \text{otherwise} \end{cases} \tag{6.14}$$

The Dykstra algorithm converges to the point in the intersection between the two sets nearest to the starting point (Bauschke and Borwein, 1994). The netSGCCA algorithm is implemented in python and accessible at `https://github.com/neurospin/netSGCCA`. It depends on the parsimony library from which it inherited the optimisation framework pylearn-parsimony package `https://github.com/neurospin/pylearn-parsimony` (de Pierrefeu et al., 2018; Lofstedt et al., 2016).

.

6.4 Application

In the previous part (see 6.3), we introduced netSGCCA, an extension of the SGCCA with an *a priori* graph. Graphs have already been used into non-multiblock statistical models as a penalty over the model parameters. Graphs were used in supervised models, such as survival models (Zhang et al., 2013) and regression models (Li and Li, 2008), and also in unsupervised models, such as matrix factorisation models (Zhu et al., 2021). However, it is still unclear how these graphs interact with the models and how they impact the feature selection. Additionally, with the availability of different graphs encoding different information, one can wonder how the choice of the graph meaningfully impacts the results obtained.

In this part, we present a study of the netSGCCA. First, using simulated data, we investigate the ability of netSGCCA to recover variables of interest in various conditions on the graph and expected solutions sides. The aim here is not to compare the different conditions but to study the behaviour of the graph penalty. Then, on the TCGA-LGG dataset, we compare the grouping effect and stability of netSGCCA while considering different graphs with different properties and from various bio-informatics sources. Finally, we investigate the relationship between selected features and the disease outcome in the same real dataset using survival prediction and pathway enrichment analysis.

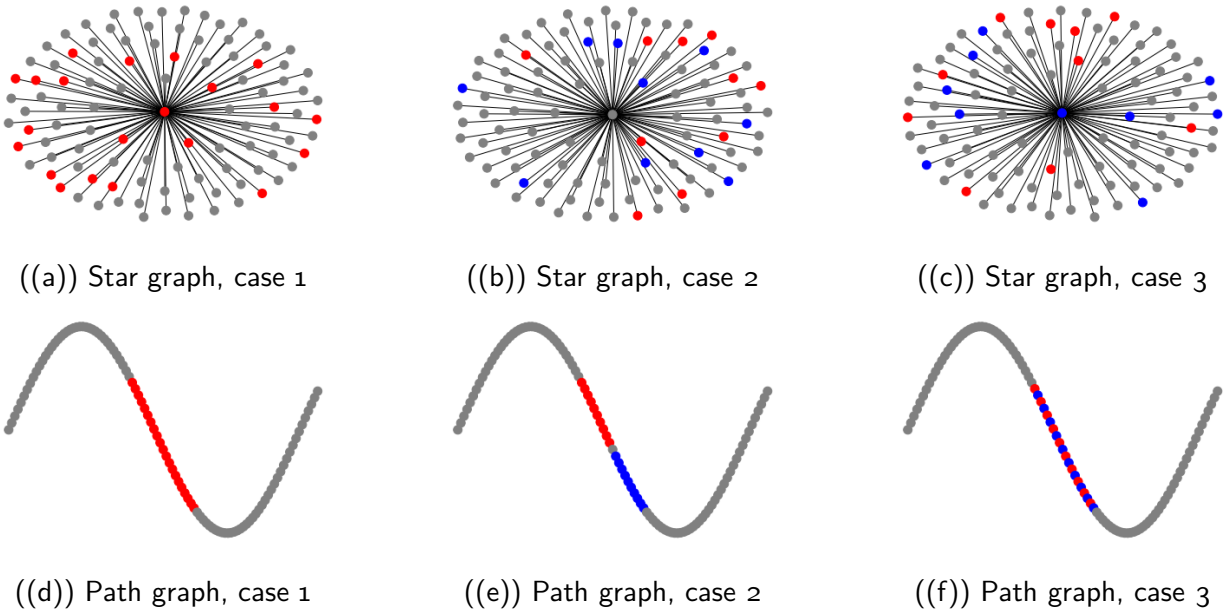
6.4.1 Simulated Data

We tested netSGCCA using two simulated blocks $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with a graph penalty based on a graph $\mathcal{G}$ on the second block. Our simulation procedure followed Du's proposal (Du et al., 2020). We started by defining the vectors $\boldsymbol{u}_1$ and $\boldsymbol{u}_2$ of dimensions $p_1 = 150$ and $p_2 = 100$. Then we generated $n = 80$ samples for each row of the two blocks $\boldsymbol{x}_1|z \sim \mathcal{N}\left(cz\boldsymbol{u}_1^\top, \Sigma_1\right)$ (respectively $\boldsymbol{x}_2|z \sim \mathcal{N}\left(cz\boldsymbol{u}_2^\top, \Sigma_2\right)$), with $z \sim \mathcal{N}(0, 1)$ a latent variable. We defined $(\Sigma_1)_{kl} = 0.1$, and $(\Sigma_2)_{kl} = -0.9 \times |u_k - u_l| + 0.9$ if the variables k and l are adjacent in the graph $\mathcal{G}$, and 0.1 otherwise. The variance of each vector is 1. c is a scalar defining the signal-to-noise ratio, in our case c takes the values 0.5 and 2.

The vector $\boldsymbol{u}_1 = (\underbrace{0,\cdots,0}_{60},\underbrace{1,\cdots,1}_{30},\underbrace{0,\cdots,0}_{60})$ was used in all configurations.

We tested 12 configurations consisting of a vector $\boldsymbol{u}_2$ and a graph. Three different cases of $\boldsymbol{u}_2$ were used to simulate different interaction types between variables of interest. The first case, $\boldsymbol{u}_2 = (\underbrace{0,\cdots,0}_{40},\underbrace{1,\cdots,1}_{20},\underbrace{0,\cdots,0}_{40})$, represents a set a variables having similar contribution to the observations and all connected to each other. The second case, $\boldsymbol{u}_2 = (\underbrace{0,\cdots,0}_{40},\underbrace{1,\cdots,1}_{10},0,\underbrace{-1,\cdots,-1}_{10},\underbrace{0,\cdots,0}_{39})$, represents two sets of variables having opposite contribution to the observations. And the third case, $\boldsymbol{u}_2 = (\underbrace{0,\cdots,0}_{40},\underbrace{1,-1,1\cdots,-1,1}_{20},\underbrace{0,\cdots,0}_{40})$, represents a two sets of variables having opposite contribution to the model and connected to each other via a non-selected variables.

Four different graphs were investigated, the path (where the edges are between subsequent variables), the star graph (where the 50th variable is connected to all the others), the union of the path and the star graph and finally, the complete graph. An illustration of the different cases can be found in Figure 6.1.



((a)) Star graph, case 1 ((b)) Star graph, case 2 ((c)) Star graph, case 3

((d)) Path graph, case 1 ((e)) Path graph, case 2 ((f)) Path graph, case 3

Figure 6.1: Star and Path graphs with different u_2 values. Grey nodes correspond to 0 values, red for 1, and blue for -1

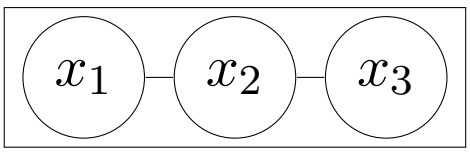


Figure 6.2: Illustration of a case were $\boldsymbol{\Sigma}_2$ does not reflect a valid correlation matrix

6.4.1.1 Practical implementation remarks

We assume that the graph edges represent interactions between the variables, which are modelled as high correlations here. The way we stated the simulation has to be further discussed in terms of numerical implementation. Suppose a graph $\mathcal{G}_1 = (\{x_1, x_2, x_3\}, \{(1,2),(2,3\})$, a graph of three variables where the second variable is connected to the first and third, as illustrated in figure 6.2. This scheme clearly states that x_1 and x_3 must be correlated. Thus, the defined $\mathbf{\Sigma}_2$ is not a plausible correlation matrix 'as is', in this case.

More generally, the defined $\mathbf{\Sigma}_2$ matrix is not guaranteed to be semi-definite. Instead, we used their nearest semi-definite matrices, relative to a weighted $\mathcal{W}$ Frobenius Norm, approximated by Higham algorithm (Higham, 2002). Having a symmetric matrix $\boldsymbol{A} \in \mathbb{R}^{p\times p}$, the Higham algorithm approximates :

$$\gamma(\boldsymbol{A}) = min\{||\boldsymbol{A} - \boldsymbol{X}|| \qquad \text{s.t} \qquad X \in \mathcal{U} \cap \mathcal{S}\} \tag{6.15}$$

With $\mathcal{S} = \{Y \in \mathbb{R}^{p\times p} \quad \text{s.t} \quad Y = Y^\top \quad \text{and} \quad Y \geq 0\}$ the set of symmetric positive semi-definite matrices. And $\mathcal{U} = \{Y \in \mathbb{R}^{p\times p} \quad \text{s.t} \quad Y = Y^\top \quad \text{and} \quad y_{ii} = 1, i = 1, \cdots, p\}$ the set of symmetric matrices with only ones in its diagonal. Higham used the Dysktra projection algorithm presented earlier 6.12.

The projection onto the set $\mathcal{U}$, relative to the $\mathcal{W}$ norm, is defined as :

$$\begin{aligned} proj_{\mathcal{U}}(A) &= A - \mathcal{W}^{-1}\text{diag}(\theta_i)\mathcal{W}^{-1} \\ \text{s.t} \quad (\mathcal{W}^{-1}\cdot\mathcal{W}^{-1})\theta &= \text{diag}(A - I) \end{aligned} \tag{6.16}$$

And the projection onto the set $\mathcal{S}$, relative to the $\mathcal{W}$ norm, is defined as :

$$proj_{\mathcal{S}}(A) = \mathcal{W}^{-\frac{1}{2}}((\mathcal{W}^{\frac{1}{2}}A\mathcal{W}^{\frac{1}{2}})_+)\mathcal{W}^{-\frac{1}{2}} \tag{6.17}$$

Such that $A_+ = Q^\top \text{diag}(max(\lambda_i, 0))Q$, where the λ_i and Q are respectively the eigenvalues and the eigenvectors of A. In our case, we chose to work with $\mathcal{W} = I$. And $\theta = diag(\boldsymbol{A} - \boldsymbol{I})$. And we refer to $\hat{\mathbf{\Sigma}}_1$ (respectively $\hat{\mathbf{\Sigma}}_2$) as the corrected correlation matrix.

6.4.1.2 Results

If we rewrite our simulation protocol, the expected covariance is :

$$\begin{aligned} z &\sim \mathcal{N}(0,1) \\ \epsilon_1 &\sim \mathcal{N}(0, I_{p_1\times p_1}) \\ \epsilon_2 &\sim \mathcal{N}(0, I_{p_2\times p_2}) \\ \boldsymbol{x}_1 &= cz\boldsymbol{u}_1^\top + \hat{\Sigma}_1^{-\frac{1}{2}}\epsilon_1 \\ \boldsymbol{x}_2 &= cz\boldsymbol{u}_2^\top + \hat{\Sigma}_2^{-\frac{1}{2}}\epsilon_2 \\ \textbf{then} \ \ \text{cov}(\boldsymbol{w}_1^\top\boldsymbol{x}_1, \boldsymbol{w}_2^\top\boldsymbol{x}_2) &= \boldsymbol{w}_1^\top\boldsymbol{u}_1\boldsymbol{u}_2^\top\boldsymbol{w}_1 \end{aligned} \tag{6.18}$$

Using the Cauchy-Schwarz, one can see that the expected covariance is maximum when $\boldsymbol{w}_1 = \boldsymbol{u}_1$ and $\boldsymbol{w}_2 = \boldsymbol{u}_2$. Thus, the model performance can be assessed on its ability to recover the variables

Table 6.1: Recovering performances depending on configurations defined by the different cases defined by the vector $\boldsymbol{u}_2$ and graphs. Corr is the correlation between the estimated components. Precision, Recall and F1 correspond to the evaluation of $\boldsymbol{u}_2$ against the computed weights. Bold refers to highest values between netSGCCA and SGCCA. Low mean to variance ratio ($c = 0.5$).

			netSGCCA				SGCCA			
		$\gamma_{\mathcal{G}}$	Corr	Precision	Recall	F1	Corr	Precision	Recall	F1
Case 1	Path	10^{-4}	$\mathbf{0.56 \pm 0.05}$	0.2 ± 0.11	0.17 ± 0.1	$\mathbf{0.18 \pm 0.1}$	0.46 ± 0.04	0.22 ± 0.32	0.03 ± 0.03	0.04 ± 0.05
	Star	1	$\mathbf{0.6 \pm 0.04}$	0.13 ± 0.07	0.3 ± 0.15	$\mathbf{0.18 \pm 0.09}$	0.48 ± 0.04	0.02 ± 0.09	0.01 ± 0.02	0.01 ± 0.04
	Union	10^{-3}	$\mathbf{0.58 \pm 0.05}$	0.13 ± 0.07	0.26 ± 0.15	$\mathbf{0.17 \pm 0.1}$	0.47 ± 0.04	0.09 ± 0.25	0.01 ± 0.03	0.02 ± 0.12
	Complete	10^{-2}	0.39 ± 0.04	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	$\mathbf{0.4 \pm 0.04}$	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
Case 2	Path	10^{-4}	$\mathbf{0.57 \pm 0.04}$	0.27 ± 0.1	0.22 ± 0.09	$\mathbf{0.24 \pm 0.09}$	0.47 ± 0.05	0.24 ± 0.32	0.04 ± 0.05	0.06 ± 0.07
	Star	1	$\mathbf{0.6 \pm 0.05}$	0.2 ± 0.09	0.48 ± 0.24	$\mathbf{0.28 \pm 0.13}$	0.47 ± 0.05	0.24 ± 0.32	0.03 ± 0.03	0.05 ± 0.06
	Union	10^{-4}	$\mathbf{0.58 \pm 0.03}$	0.21 ± 0.08	0.43 ± 0.15	$\mathbf{0.29 \pm 0.1}$	0.47 ± 0.04	0.32 ± 0.39	0.04 ± 0.06	0.07 ± 0.1
	Complete	10^{-1}	0.35 ± 0.04	0.01 ± 0.03	0.01 ± 0.06	$\mathbf{0.01 \pm 0.04}$	$\mathbf{0.38 \pm 0.04}$	0.01 ± 0.03	0.0 ± 0.01	0.0 ± 0.02
Case 3	Path	10^{-2}	$\mathbf{0.56 \pm 0.04}$	0.16 ± 0.07	0.3 ± 0.14	$\mathbf{0.21 \pm 0.1}$	0.45 ± 0.03	0.1 ± 0.25	0.01 ± 0.02	0.02 ± 0.04
	Star	1	$\mathbf{0.63 \pm 0.05}$	0.17 ± 0.06	0.37 ± 0.11	$\mathbf{0.24 \pm 0.07}$	0.49 ± 0.04	0.12 ± 0.15	0.02 ± 0.03	0.03 ± 0.04
	Union	10^{-4}	$\mathbf{0.59 \pm 0.03}$	0.2 ± 0.05	0.39 ± 0.11	$\mathbf{0.26 \pm 0.07}$	0.46 ± 0.04	0.16 ± 0.28	0.02 ± 0.03	0.04 ± 0.06
	Complete	10^{-2}	0.39 ± 0.05	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	0.39 ± 0.05	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0

Table 6.2: Recovering performances depending on configurations defined by the different cases defined by the vector $\boldsymbol{u}_2$ and graphs. Corr is the correlation between the estimated components. Precision, Recall and F1 correspond to the evaluation of $\boldsymbol{u}_2$ against the computed weights. Bold refers to highest values between netSGCCA and SGCCA. High mean to variance ratio ($c = 2$).

			netSGCCA				SGCCA			
		$\gamma_{\mathcal{G}}$	Corr	Precision	Recall	F1	Corr	Precision	Recall	F1
Case 1	Path	10^{-3}	0.73 ± 0.04	1.0 ± 0.0	1.0 ± 0.0	$\mathbf{1.0 \pm 0.0}$	$\mathbf{0.76 \pm 0.04}$	1.0 ± 0.0	0.2 ± 0.05	0.32 ± 0.06
	Star	10^{-4}	0.73 ± 0.05	0.46 ± 0.03	1.0 ± 0.0	$\mathbf{0.63 \pm 0.03}$	$\mathbf{0.74 \pm 0.03}$	1.0 ± 0.0	0.21 ± 0.06	0.35 ± 0.07
	Union	10^{-4}	0.72 ± 0.04	0.47 ± 0.05	1.0 ± 0.0	$\mathbf{0.64 \pm 0.04}$	$\mathbf{0.74 \pm 0.04}$	1.0 ± 0.0	0.21 ± 0.06	0.35 ± 0.08
	Complete	10^{-4}	0.36 ± 0.09	0.02 ± 0.07	0.06 ± 0.23	0.03 ± 0.11	$\mathbf{0.39 \pm 0.09}$	0.05 ± 0.22	0.02 ± 0.09	0.03 ± 0.13
Case 2	Path	10^{-3}	0.71 ± 0.04	1.0 ± 0.0	1.0 ± 0.0	$\mathbf{1.0 \pm 0.0}$	$\mathbf{0.74 \pm 0.04}$	1.0 ± 0.0	0.24 ± 0.08	0.38 ± 0.11
	Star	10^{-4}	0.7 ± 0.04	0.47 ± 0.03	1.0 ± 0.0	$\mathbf{0.64 \pm 0.03}$	$\mathbf{0.74 \pm 0.04}$	1.0 ± 0.0	0.19 ± 0.06	0.32 ± 0.09
	Union	10^{-4}	0.71 ± 0.04	0.47 ± 0.02	1.0 ± 0.0	$\mathbf{0.64 \pm 0.02}$	$\mathbf{0.74 \pm 0.04}$	1.0 ± 0.0	0.22 ± 0.05	0.36 ± 0.07
	Complete	1	0.32 ± 0.04	0.02 ± 0.1	0.05 ± 0.22	0.03 ± 0.13	$\mathbf{0.39 \pm 0.09}$	0.05 ± 0.22	0.02 ± 0.09	0.03 ± 0.13
Case 3	Path	10^{-3}	0.73 ± 0.04	1.0 ± 0.0	1.0 ± 0.0	$\mathbf{1.0 \pm 0.0}$	$\mathbf{0.74 \pm 0.04}$	1.0 ± 0.0	0.21 ± 0.06	0.34 ± 0.08
	Star	10^{-4}	0.7 ± 0.05	0.47 ± 0.04	1.0 ± 0.0	$\mathbf{0.64 \pm 0.04}$	$\mathbf{0.74 \pm 0.05}$	1.0 ± 0.0	0.2 ± 0.06	0.33 ± 0.08
	Union	10^{-4}	0.71 ± 0.04	0.46 ± 0.04	1.0 ± 0.0	$\mathbf{0.63 \pm 0.04}$	$\mathbf{0.75 \pm 0.03}$	1.0 ± 0.0	0.19 ± 0.07	0.31 ± 0.1
	Complete	10^{-4}	0.36 ± 0.09	0.02 ± 0.07	0.07 ± 0.23	$\mathbf{0.03 \pm 0.11}$	$\mathbf{0.39 \pm 0.09}$	0.05 ± 0.22	0.01 ± 0.07	0.02 ± 0.1

of interest and maximise the covariance between extracted latent variables. To do so, we computed the precision, recall and F1 metrics between the true $\boldsymbol{u}_2$ vectors and the weights $\boldsymbol{w}_j$ estimated by the model. For each configuration, we chose the hyper-parameter $\gamma_{\mathcal{G}}$ by running the model 20 times, with $\gamma_{\mathcal{G}}$ ranging from 10^{-4} to 10^4 each time. The best $\gamma_{\mathcal{G}}$ was selected using the best average F1 score because our objective is mainly to recover the variables of interest. We also ran the model for each configuration without using a graph and compared the results. The sparsity value was fixed to $\sqrt{25}$ for the first block (resp. $\sqrt{20}$ for the second block) in all runs.

Table 6.1 and Table 6.2 show the results obtained on the simulated data. It shows that when we have a high mean/variance ratio, the F1 scores are higher, which means that the models using the graphs focus on retrieving the underlying projector $\boldsymbol{u}_2$. This was expected since a low mean/variance ratio means noisier data.

Overall, the tables also show that using netSGCCA outperformed the SGCCA without a graph. When $c = 2$ (mean to variance ratio), SGCCA selected very few variables, about 3, leading to a high precision but very low recall. In contrast, the graph penalisation allowed the model to select more variables, retrieving all the variables of interest and a high F1 score. However, this increase

of the F1 score came with a slight decrease in the correlation between the estimated components, by around 2%. Additionally, when $c = 0.5$, the F1 score continues to show improvement when the graph penalisation is used, but to a lower degree. However, the correlation between the estimated components also increased by 0.13 on average.

Comparing each graph type, we can see that, when $c = 2$, the path graph recovers all the variables of interest perfectly, with an average F1 score of 1 in all cases. The star graph was also able to obtain a perfect recall but with much lower precision. This is because the star graph selects many more variables, about 45 on average. Knowing that the sparsity level is the same for all configurations, the hub in the star graph seems to spread the weights more into its neighbours compared to the path graph. However, the correlations between estimated components are comparable. When $c = 0.5$, the models seem to select the variables randomly, which is shown by an F1 score close to 0.2. Additionally, the weights do not seem to resemble the original $\boldsymbol{u}_2$. However, even this result is better than without the graph *a priori*, which only selected a couple of features and resulted in an F1 score close to 0. In this situation, by choosing a greater number of variables, the star graph performed better in the correlation score compared to the path graph. Additionally, the union of the star and path graph exhibited behaviour similar to the ones of path and star graphs. Finally, the models with the complete graph always failed to outperform all the other models. This result is expected since the complete graph does not contribute to bring any information.

If we fix the graph and the mean-variance ratio, for all the cases $\boldsymbol{u}_2$ considered, we observe no significant difference in the precision of the variable selection process nor in the extracted correlations. This observation holds for all graph types and mean-to-variance ratios. The correlations between neighbours in the graph did not change the selected variables.

Overall, netSGCCA seemed to outperform the SGCCA in terms of retrieving the variables of interest, in nearly all configurations tested. It appeared that it is through its properties and structures that the graph have an influence on the behaviour of the model. We seek to investigate these results on real oncological data in the next sections.

6.4.2 TCGA-LGG analysis design and results

Table 6.3: Different sources of prior knowledge graphs

	# nodes	# edges	Diameter	Radius	% isolated nodes
PC	15710	841690	6	4	0.20
MSIGDB	10463	82962	6	4	0.46
KEGG	776	11963	12	7	0.96

To assess the behaviour of the proposed method on real data, we used the TCGA-LGG dataset. The dataset includes 419 patients. All patients have CNV, which has 57964 variables, miRNA with 645 variables, and RNA with 22297 variables.

netSGCCA (and SGCCA) is an unsupervised model that jointly reduces the dimension of the different blocks. We want learned features that not only maximise the correlation between different blocks but are also relevant to survival. Segal (2005) previously demonstrated that solving a sparse Cox model is equivalent to solving a linear regression when the null deviance residuals replace the

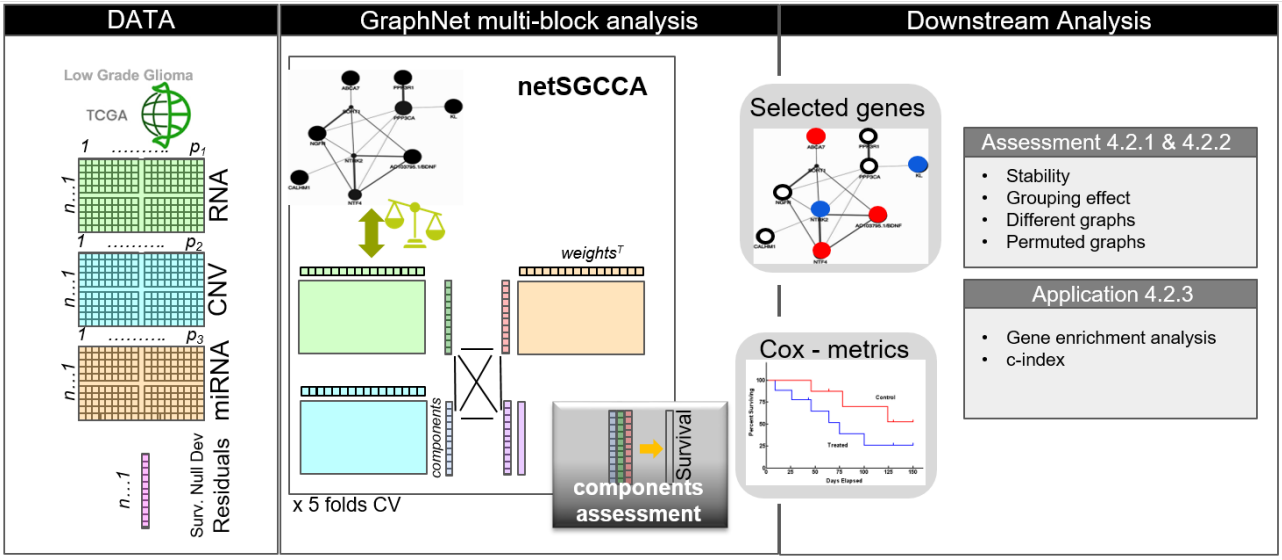


Figure 6.3: Details of the proposed experimental design for the TCGA-LGG dataset. The analysis step refers to the comparison between the different graphs, analysis of the variable selection when nodes are permuted or edges are removed. Application refers to survival prediction and enrichment analysis.

survival. The null deviance residuals are derived from the highly skewed martingale residuals by a normalising transformation analogous to deviance residuals for Poisson regression :

$$\begin{aligned} D_i &= \text{sign}(\hat{M}_i)\sqrt{-\hat{M}_i - \delta_i \log((\delta_i - \hat{M}_i)/\delta_i)} \\ &\approx \frac{\delta_i - \hat{E}_i}{\sqrt{\hat{E}_i}} \end{aligned} \tag{6.19}$$

Where δ_i indicates whether or not an event is observed for subject i, $\hat{E}_i$ is the expected number of events at time t_i (event time for individual i), and $\hat{M}_i = \delta_i - \hat{E}_i$ is the martingale residuals.

Using this equivalency, Bastien (2008) proposed to predict survival using the PLS regression model by replacing the outcome with the null deviance residuals. Similarly, we added the null deviance residuals into the netSGCCA (and later to SGCCA). We aim to make the model learn to correlate latent features extracted from the other blocks to survival. Note that this extra block is added only during training and does not intervene during the inference stage.

We applied the GraphNet penalty on the RNA block. Gene identifiers were mapped to their HUman Genome Organisation (HUGO) names with non-matching genes removed, leading to 19864 genes remaining in the RNA block.

Different gene-gene interaction graphs with different properties were used. Pathway Commons v12 (PC) (Cerami et al., 2010) is an aggregation of multiple subgraphs from various sources, containing 15710 nodes and 841690 edges. The Molecular Signature Data Base (MSIGDB) (Liberzon et al., 2011; Subramanian et al., 2005) C3 regulatory target gene set is one of the subgraphs of the PC graph containing 10463 nodes and 82962 edges. The Kyoto Encyclopedia of Genes and Genome (KEGG) (Kanehisa, 2000) is also another subgraph of the PC graph, containing 776 nodes and 11963 edges. The full description of these graphs can be found in the Table 6.3. Since graphs do not include all genes, missing genes were added as isolated nodes, and genes mapped to the same HUGO name were duplicated in the graph with their edges.

We aimed to compare the gene selection abilities of the different graphs. First, using the PC graph, we established the differences between the raw and normalised Laplacian, in terms of the number of selected variables, as the $\gamma_{\mathcal{G}}$ varies, and the stability of variable selection. Then, we compared the different graphs. Later, to establish the impact of the graph in the selection process and disentangle the graph penalty effect from the RGCCA model effect, we permuted the nodes of the PC graph and compared the results with the non-modified graph. We also removed edges from the PC graph to investigate the effect of the graph density and the importance of the edges between selected nodes. Finally, we evaluated the application of the features extracted by using them to predict survival and discussed the biological pathways potentially involved through the selected genes. Details of the experimental design are presented in Figure 6.3.

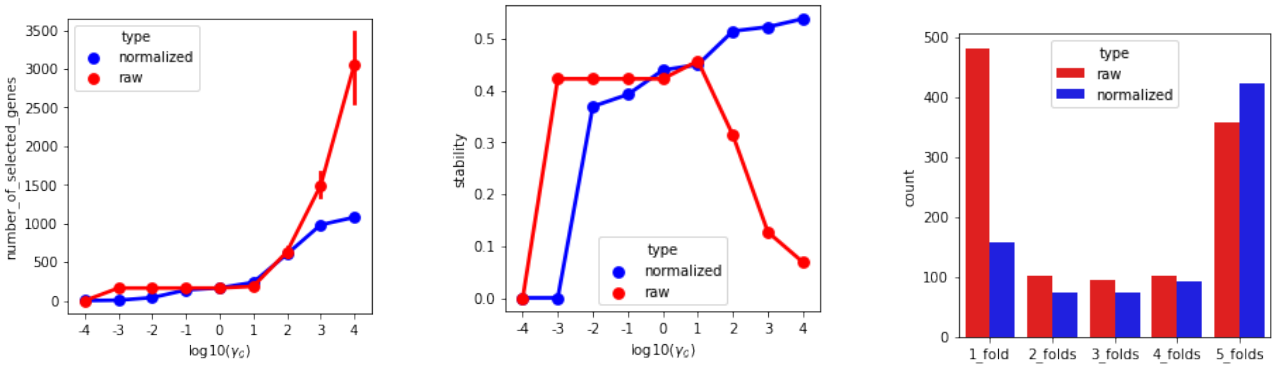
In order to achieve the experimental design, we first isolated 15% of the patients (63 patients) as a test set and performed the training on the remaining 356 patients. We stratified the patients using the event status. The training set was split into five folds of equal sizes. This allowed us to compute a standard deviation for the c-index metric, and identify two sets of selected genes: the **candidate selected genes** (the union of gene sets selected in each fold) and **stable selected genes** (the intersection of gene sets selected in each fold). The test set was only used for the survival prediction in the final part of our study.

Throughout this work, $c_{j,k} = 1$ for all j and k, with $j \neq k$. This means that each block is connected to all the other blocks in the netSGCCA model. Additionally, the ℓ_1 constraints have been fixed as the best yielding parameters - in terms of the c-index - on the five folds without GraphNet using a grid search.

6.4.2.1 Comparison between normalised and raw Laplacian

To compare the normalised Laplacian with its raw version, using the PC graph, we varied the $\gamma_{\mathcal{G}}$ between 10^{-4} and 10^{4}. Figure 6.4(a) shows the evolution of the number of selected genes, for both Laplacians, as $\gamma_{\mathcal{G}}$ increases. We can see that the higher the $\gamma_{\mathcal{G}}$, the more genes are selected; this is in line with previous findings in the related works. It also shows that, without a graph penalty, very few genes are selected (about 3 in each fold). The Figure 6.4(a) also highlights a similar behaviour between the raw Laplacian and the normalised Laplacian, in terms of the number of genes selected. However, a closer inspection shows that the normalised Laplacian is more stable as $\gamma_{\mathcal{G}}$ increases. To show this, we defined the stability as the $\frac{\#\,of\ stable\ selected\ genes}{\#\,of\ candidate\ selected\ genes}$ and computed it for each $\gamma_{\mathcal{G}}$. Results are shown in Figure 6.4(b). Figure 6.4(c) shows the distribution of the number of folds in which a gene was selected, when $\gamma_{\mathcal{G}}$ was fixed to 10^2. In the case of the normalised Laplacian, around 51% of genes selected by the model were chosen in all five folds. In contrast, only 30% were selected in all five folds with the raw Laplacian. In fact, a large number of genes were selected only once.

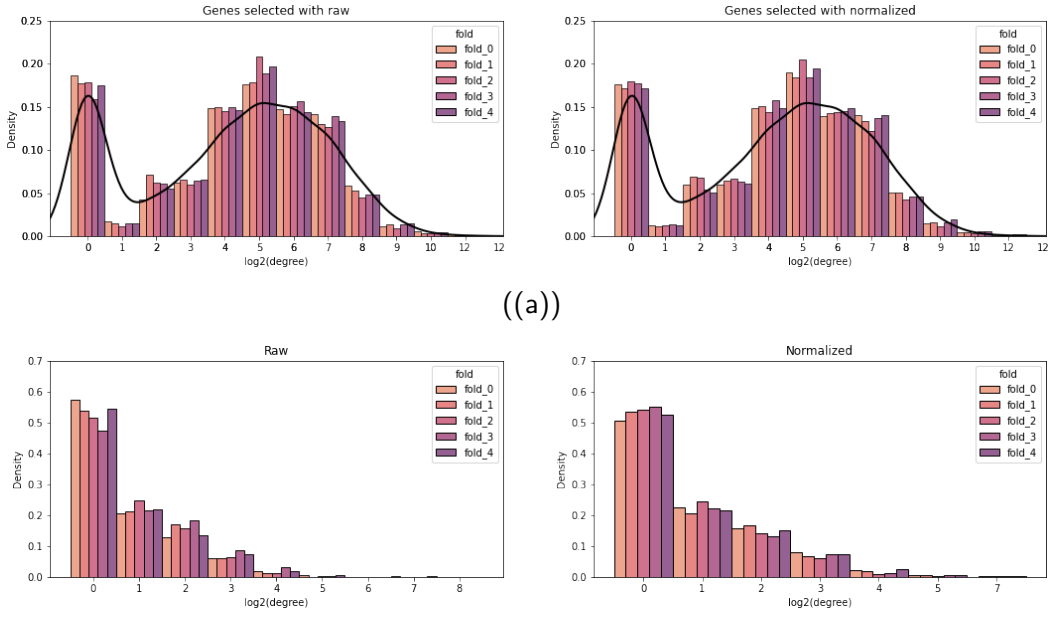
We examined the degree distributions by considering the subset of selected genes across five folds in the full PC graph on the one hand and, on the other hand, in the PC sub-graph containing the selected genes. Figure 6.5(a) shows the degree distribution of selected genes in the full PC graph. Both the normalised and raw Laplacian follow the same behaviour. The two distributions are also similar to the degree distribution considering all genes in the PC graph. Considering the proportion of isolated nodes in PC, they are underrepresented among the selected genes. This indicates that



((a)) ((b)) ((c))

Figure 6.4: Comparison between Raw and Normalised Laplacian. (a) Evolution of the number of selected genes as $\gamma_{\mathcal{G}}$ varies, using the raw and normalised Laplacian. (b) Evolution of the stability metric as $\gamma_{\mathcal{G}}$ varies, using the raw and normalised Laplacian. (c) Number of selection co-occurrence of selected genes, with $\gamma_{\mathcal{G}} = 10^2$.

GraphNet does not favour genes with a high degree in the graph but discriminates against isolated nodes. However, GraphNet did not select neighbours in the graph. This is exhibited in Figure 6.5(b), which gives the degree distribution in the PC sub-graph of the selected genes. It shows a shift toward lesser degrees suggesting that genes are mostly selected because of their covariance similarity across the patients, but not because of their neighbourhood in the graph penalty (also shown later). The GraphNet penalty, in our configuration, seems to mainly smooth weights over similarly correlated genes but does not give rise to gene communities because of their adjacency in penalty graph.



((a))

((b))

Figure 6.5: Degree distribution of selected genes by fold for Raw and Normalised Laplacians. (a) For each selected gene, we counted the number of its neighbours in the PC graph. The black line represents the density of the degree distribution of all genes in the PC graph. (b) For each selected gene, we counted the number of its neighbours among selected genes.

We looked at netSGCCA gene weight distribution according to the degree of the gene in the PC graph. Figure 6.6 shows that the gene degree does not influence the weight. However, isolated nodes tend to have significantly higher weight variance, ranging from -0.1 to 0.1 for the normalised Laplacian. This is expected as the more a node is connected, the more its weight is constrained. The same pattern can be shown for the normalised and raw Laplacian, except that the normalised Laplacian produced smaller weights in absolute value. Additionally, Figure 6.7 shows the weight difference between genes according to the distance between the genes in the graph (we used the shortest path in the graph). It shows that the closer the genes, the closer their final weights, going from 0.016 on average between neighbouring nodes to 0.017 if the shortest path between them is five, for the normalised Laplacian. The same pattern can be seen on the raw Laplacian with higher values. Thus, even if the graph does not select sub-networks, it has a grouping effect.

Finally, the model seems to select genes by correlation, which is exhibited in Figure 6.8 which shows correlation distribution among selected genes by the normalised Laplacian. The model selected two distinct, negatively correlated groups. The correlation distribution between selected genes does not follow the distribution of the correlation among gene expression profiles in the whole LGG dataset. The GraphNet penalty seems to partially override the ℓ_1 penalty and allow the model to select groups of highly correlated variables.

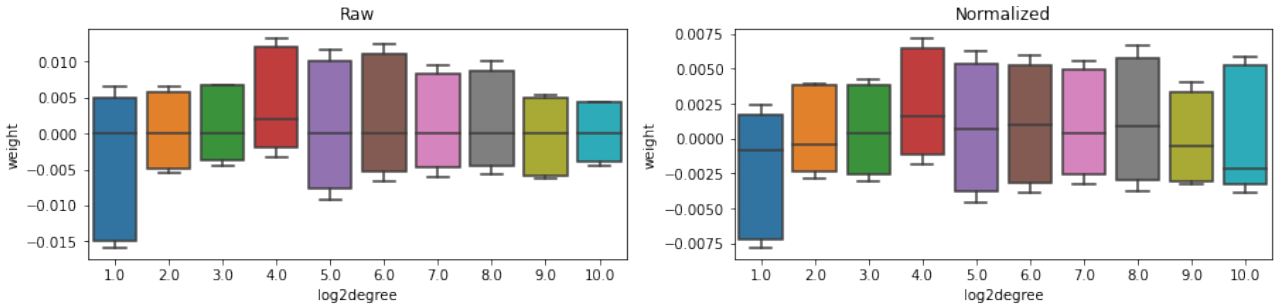


Figure 6.6: Box plot of the weight distribution of selected genes for Raw and Normalised Laplacians. x-axis is log2 of the gene degrees in full PC graph.

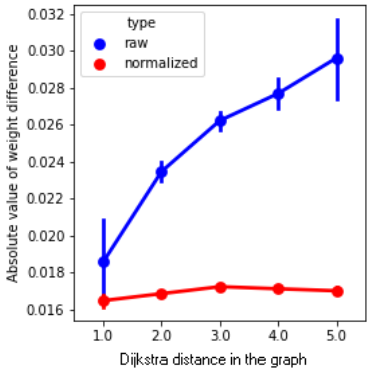


Figure 6.7: Average absolute weight difference between selected nodes according the Dijkstra distance between them in the PC graph

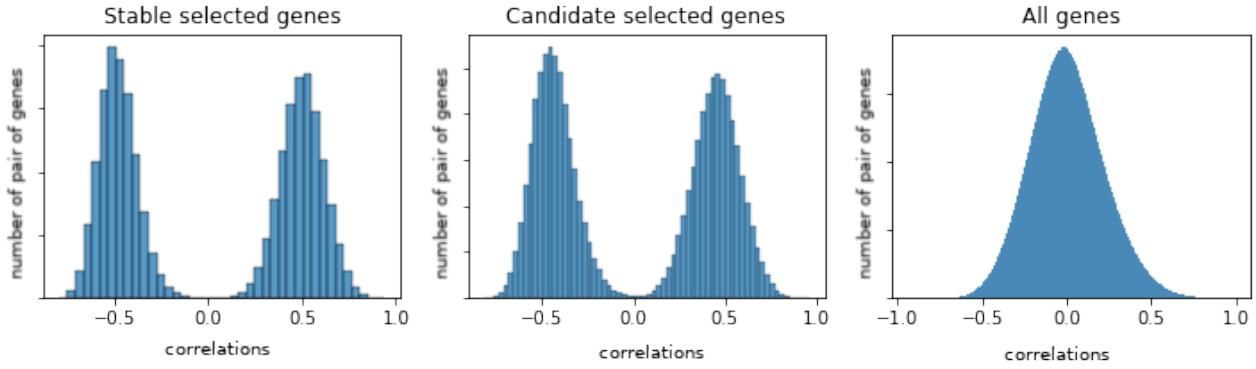


Figure 6.8: Correlation distribution of selected genes. Between stable selected genes (Intersection of selected genes in the 5 folds), Candidate selected genes (Union of selected genes in the 5 folds) and All the genes in the dataset.

6.4.2.2 Comparisons between different graphs

Since the normalised Laplacian allowed the selection of more stable gene sets, the following experiments will use only normalised Laplacians. We ran our model using the MSIGDB and KEGG graphs on the same five-folds and the same model hyper-parameters. We compared the number of genes selected by each graph and the overlap of selected genes between each graph and netSGCCA using the PC normalised Laplacian. More precisely, we compared the overlap between the stable selected genes and the candidate selected genes. Figure 6.9 shows the results obtained using the MSIGDB and KEGG Laplacian. Both graphs selected fewer genes than the PC graph, with the MSIGDB selecting more than the KEGG. However, there is a large overlap between the selected genes given each graph penalty. This suggests that the graph structure strongly influences the grouping effect of the method, but also suggests that it does not substantially impact the set of genes selected.

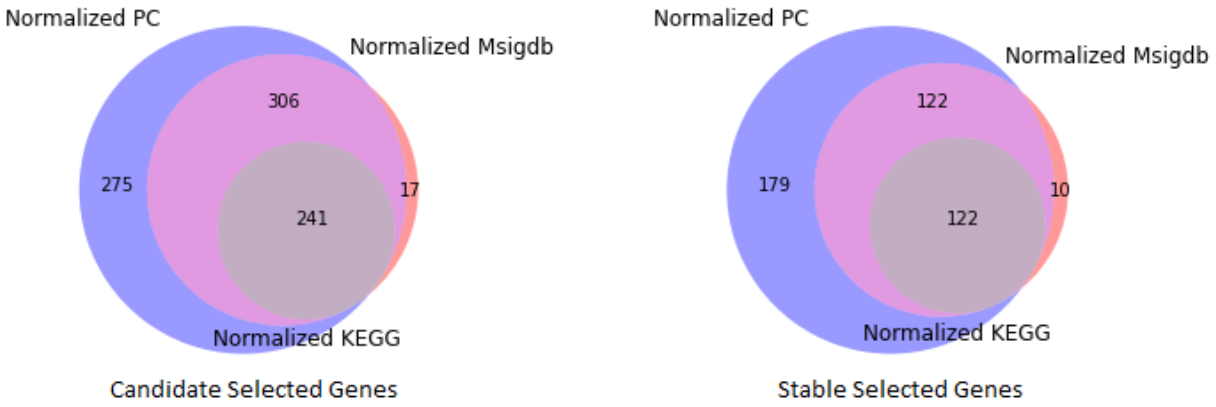


Figure 6.9: Venn diagram showing the overlap between genes selected by the PC graph and the MSIGDB and the KEGG Graphs. Diagrams on the right shows the gene sets resulting from the candidate selected genes, while left diagrams show the gene sets from the stable selected genes

To better exhibit the influence of the graph semantic (information embedded in the graph), we randomly permuted gene labels within the PC graph, and using this new graph, we extracted genes in terms of stable or candidate genes. This was done ten times to check the validity of the results.

For the stable and candidate sets, we computed the Intersection Over Union metric (IOU) from each permutation and the selected genes using the PC graph. The different permuted graphs selected a similar number of candidate selected genes as the original PC graph, but the IOU was always below 75% as shown in Figure 6.10. Same applies for the stable selected genes. This illustrates that, for a biological graph of a given density, its semantics is weakly passed by the GraphNet penalty and has a limited influence on the selected genes.

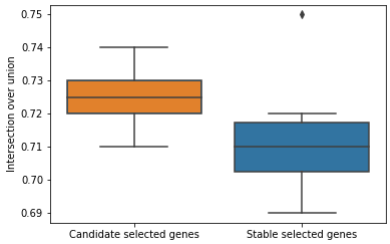


Figure 6.10: Box plot for the IOU metric between the genes selected by the PC graph and the permuted PC graph

Table 6.4: The effect of pruning edges that connect genes selected when using the full PC graph. Candidate and stable genes sets obtained when using the different pruned sub-graphs.

		Ref. set is full PC graph Candidate (822)	Ref. set is full PC graph Stable (423)
	Edges removed in the pruned subgraphs	# of selected genes	# of selected genes
Candidate	Inner	822	423
	Outer	586	262
	All	238	107
Stable	Inner	822	423
	Outer	450	220
	All	239	108

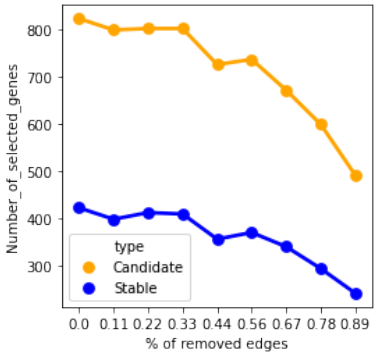


Figure 6.11: The evolution of the number of selected genes when the number of edges decreases (x-axis is the percentage of edges removed from 0% to 90%.)

To show the impact of the graph density on the number of selected genes, we removed edges randomly from the PC graph. We made sure that we removed edges from candidate selected genes

using the normalised PC graph in the same proportions as from the whole edge set. Figure 6.11 shows the evolution of the number of selected genes as edges are removed. It exhibits a strong correlation between graph density ($\frac{\#\text{ of edges}}{\#\text{ of all possible edges}}$) and the number of selected genes. This is the case for both candidate and stable selected genes. Additionally, selected genes after the removal of edges are all included in the original set of selected genes, for both candidate and stable. This shows that GraphNet penalty achievement for feature selection depends on the density of the graph it used.

We investigated the importance of the direct and the indirect paths between the genes selected by a model with the GraphNet penalty based on the full PC graph. To do this, we removed inner edges (edges connecting two selected genes), outer edges (edges connecting a selected and non-selected gene), and all edges of the selected nodes, which would make them completely isolated. We considered the candidate and stable selected gene sets obtained with the PC normalised Laplacian, which produced 822 candidate selected genes from the five folds of which 423 were stable selected genes. As shown in Table 6.4, removing the inner edges (around 1450 edges) did not change either the number nor the set of genes selected. Removing the outer edges (around 65285 edges) diminished the number of selected genes while keeping a large overlap. Making all the originally selected genes isolated nodes reduce the number of selected genes by more than a half. Even when removing either the inner or the outer edges, some path may remain between the selected genes, which helps the model to retrieve them. However, making them isolated lowers their probability of being selected.

6.4.2.3 Survival Prediction

Table 6.5: Performances in survival prediction, number of selected variables and pathways depending on the type of graph to constrain the model.

	C-index		Stable variables selected in block			No. of pathways from the RNA
	Validation set	Test set	RNA	MiRNA	CNV	
no graph	0.708 ± 0.122	0.588	0	1	208	0
raw PC	0.692 ± 0.127	0.632	357	0	0	3
normalised PC	0.709 ± 0.141	0.589	423	1	0	3
normalised MsigDB	0.741 ± 0.092	0.625	232	1	0	2
normalised KEGG	0.712 ± 0.110	0.690	106	1	0	2

In the previous part, we presented results regarding the variable selected in the various blocks of our multi-block setting, but netSGCCA also yields components for each blocks. Using the components extracted from the CNV, RNA, and the miRNA by the netSGCCA method, we performed a survival prediction using a simple Cox model. We compared the results using the different graphs and a model without a graph constraint, and we present the results in Table 6.5. It shows that models using the GraphNet penalty have similar C-index scores compared to the model without a graph. When using the MSIGDB and KEGG, results are slightly better compared to the other models on the validation and test sets. This indicates that the GraphNet penalty did not weaken the ability of the estimated components to predict patients' survival. However, the model without GraphNet could not reliably extract genes, which hindered its ability to identify pathways of interest. SGCCA, without a graph, selected about three variables per fold. It is in line with the results found so far, as the ℓ_1 penalty tends to select few representatives among highly correlated variables. GraphNet forced the model to gather

these correlated variables. Using the different graphs for the GraphNet penalty did not change the C-index noticeably, which is also expected as the variable selection has shown a substantial overlap. Adding the graph made the signatures from other blocks less stable. The more stable RNA variables are selected, and the fewer and less stable variables are observed from the CNV block. As a result, no stable variables are obtained from the CNV block when a graph was used, but it must be balanced by the fact that for each fold, the model selected about 150 variables when the normalised Laplacian was used (compared to 250 when no graph was used).

Table 6.6: Top gene sets from MSigDB C6 collection. In bold, gene sets with an adjusted p-value lower than 0.05.

Term	Adjusted P-value
CAHOY ASTROGLIAL	0.000384
ATF2 UP.V1 DN	0.016688
TGBF UP.V1 DN	0.041445
KRAS.KIDNEY UP.V1 UP	0.104489
RAF UP.V1 DN	0.260525

Genes function together in biological pathways. In order to find the most significantly represented pathways by the selected genes, we used Enrichr (Chen et al., 2013; Kuleshov et al., 2016; Xie et al., 2021). Enrichr uses the Fisher exact test to compute the significance of the enrichment of a pathway. This is a proportion test that assumes a binomial distribution and independence for the probability of any gene belonging to any set. We used the MSigDB C6 collection to investigate associations between signatures and the 423 selected genes.

Results are presented in Table 6.6. We found pathways that have already been associated with low-grade gliomas. Genes associated with astrocytes in the set CAHOY ASTROGLIAL, have also been selected by the model. These genes have previously been studied for their link with brain tumour development (Irvin et al., 2017; Katz et al., 2012). *ATF2* is known to promote invasion in malignant glioma (Zhang et al., 2014). Additionally, *TGF-β* (Transforming growth factor-beta) has been targeted to limit brain tumour growth (Han et al., 2015). Other pathways have been labelled by enrichment analysis but the corrected p-value was not significant. For example, the *RAF* Fusion has been associated with pediatric low-grade tumours (Lind et al., 2021). Mutations in *KRAS*, *HRAS* and *NRAS* are known in gliomas and are often concomitant with *BRAF* mutations and fusions (Knobbe et al., 2004).

6.4.3 Discussion

Our work establishes some characteristics of netSGCCA, a data integration method that implements the GraphNet penalty. It shows, in the context of multi-view analysis, that GraphNet helps to group variables instead of selecting a few candidates, which is in line with previous results involving the same penalty. We also exhibit better interpretation capabilities for netSGCCA compared to SGCCA, as it allows the selection of sound and stable candidate variables within a block. When netSGCCA implements the GraphNet penalty using the normalised Laplacian of the *a priori* graph and not the raw, it leads to an even greater stability.

From the application of netSGCCA to a real multi-modal oncological dataset, we have derived presumably general observations. For the block being submitted to the GraphNet penalty, the similarity of the variable profiles is the primary driver for co-selecting variables, and the proximity of the variables (as nodes) in the graph is secondary. The graph seems to present candidates that are grouped according to their correlations. The density of the graph *a priori* used in the GraphNet strongly influences the final number of selected variables. The denser the graph, the more variables were selected. Yet, our results also suggest that the capacity of the netSGCCA at extracting variables of interest capitalises mainly on the variables initially selected by the SGCCA (without a graph). Finally, the large overlap between selected variables in the penalised block when using graphs with equivalent density but with different semantics, suggests the netSGCCA model relies on the data first.

Regarding the LGG pathology, as it may be studied from the TCGA multi-modal dataset, we demonstrate two remarkable achievements of netSGCCA that overrides the current performance of other multi-modal integration frameworks. First, netSGCCA predicts survival very well. Large c-index values were obtained without using the medical data (no eCRF information other than survival). The results found are similar to reference values reported in a recent work which did consider medical data (Herrmann et al., 2020). Second, using netSGCCA with the PathwayCommon graph penalty for the gene expression block, we took advantage of the stability of variable selection to propose a list of candidate genes and candidate biological pathways that explain the pathological outcome. Overall, this suggests that graph penalty in multi-modal analysis model like netSGCCA is able to bring original molecular biology insights into the pathology.

A limitation of this study is the small number of applications considered, namely one simulation and the TCGA LGG dataset. While a general observation is that GraphNet selects more variables than the classical Elastic-Net, discussions remain about its stability when there are multiple variables of interest with no correlation between them. It is a data-related problem, and the model performance on the LGG dataset is insufficient to give indications about its behaviour in such a case.

6.4.4 Conclusion

The present work focuses on the analysis of the GraphNet penalty available in the mulit-block netSGCCA model and applied to the TCGA-LGG dataset. Contrary to Elastic-Net alone, GraphNet penalty is able to select a reasonable set of genes and yields informative biological interpretation from the pathway enrichment analysis. The example on the TCGA-LGG dataset exhibits the stability and reliability of netSGCCA for selecting variables of interest. However, it is important to note that we show that the co-selection of variables is not primarily influenced by the structure of the graph, but rather by its overall density. Therefore, an interpretation in terms of the paths read in the graph is elusive. Nevertheless, the method did extract genes that have been found co-(de)regulated in other studies of low-grade gliomas and other brain tumours. Future applications should focus on extending the results to other tumour types. Additionally, the multi-block model should enlarge the scope of multi-modality beyond molecular data. New data sources can be investigated, such as imaging data with their specific penalty resources, in order to increase prediction performance or unveil shared

information between modalities.

* * *
* *
*

Chapter 7

Radio-genomics integration and survival prediction

Radiomic features have been successfully used for different tasks and in various tumour types. This includes the DIPG, Glioblastomas and Low-Grade Gliomas (LGG). However, it is unclear which image modality is the most informative or which feature type is the most important. Furthermore, research has also shown that adding radiomic features to molecular data can improve survival prediction. But it has yet to be integrated into a full multi-block framework. This should help us understand the interactions between the radiomic and genomic features.

So far, we have proposed a procedure to extract the regions of interest required for the radiomic analysis. Additionally, we proposed the netSGCCA, a multi-block model that considers an *a priori* graph of interactions. We also established basic properties that help us understand the interaction between the multi-block model and the graph penalty.

The current chapter is devoted to applying netSGCCA for survival prediction when genetic data are associated with radiomic features. We pay close attention to the selected variables and their link to the studied disease. Since the multi-block framework we are working with requires data availability for all sources on all samples, it naturally results in fewer samples being analysed. Therefore, we establish the baseline performance of each block when all available data is considered before the integration. Then, we compare the mono-block framework with the multi-block framework. We also compare the netSGCCA with the best-performing models in the related works. Due to the unavailability of survival data on the DIPG dataset, this study was restricted to the TCGA-LGG dataset.

Chapter Outline

Contents

7.1 Methodology . . . 93
7.1.1 Other Multi-block survival models from the state-of-the-art . . . 93
7.1.2 Experimental design . . . 94
7.2 Baseline results using all per-block available data . . . 97
7.2.1 Radiomics . . . 97
7.2.2 Somatic mutations . . . 100
7.2.3 RNA . . . 103
7.3 Comparison mono-block and multi-block approaches . . . 104
7.3.1 Mono-block . . . 104
7.3.2 Multi-block . . . 108
7.4 Discusion . . . 111
7.5 Conclusion . . . 112

7.1 Methodology

7.1.1 Other Multi-block survival models from the state-of-the-art

Our strategy using netSGCCA and SGCCA is similar to the one introduced in the previous chapter. We used joint dimensionality reduction, followed by a Cox model. In this strategy, we consider the null *Cox deviance residuals* as an extra block, as proposed by Bastien (2008). This allows taking into account survival information in the unsupervised dimensionality reduction. We note that this survival block is used only during unsupervised training and not by the Cox model.

This section presents different state-of-the-art multi-block models, which will be compared to netSGCCA and SGCCA in our survival analysis. All used models aim at maximising the log-likelihood function presented in equation 2.7. We compared netSGCCA and SGCCA with multi-block supervised survival models, which directly infer the risk of each patient. We consider $\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(J)}$ of sizes $n \times p^{(1)}, \cdots, n \times p^{(J)}$ the different blocks of data obtained on the same set of patients.

We tested two supervised models, Priority-Lasso (Klau et al., 2018) (see also 3.3.2.1) and BlockForest (Hornung and Wright, 2019) (see also 3.3.2.2), whose features will be detailed below. We chose these two algorithms because, in Herrmann's benchmark study (Herrmann et al., 2020), these two methods obtained the highest c-index among the compared methods. We note that, even if BlockForest does not perform variable selection, we added it to our study as it presented the best results on the various cancers benchmarked.

Priority-Lasso Klau et al. (2018) proposed a multi-block adaptation of a linear model, which includes the Cox model. The algorithm requires a priority order of the blocks. This order is user-chosen and may reflect hypotheses about the blocks performances, or their importance to the issue at hand.

Let $\boldsymbol{X}^{(1)}, \cdots, \boldsymbol{X}^{(J)}$ be the different blocks of data. Here we suppose that the blocks are already ordered, with $\boldsymbol{X}^{(i)}$ having a higher priority than $\boldsymbol{X}^{(i+1)}$. The algorithm sequentially fits a model on each block and uses the results of the previously fitted model as an offset. Since the outcome $\boldsymbol{Y}$ is a part of the data used to estimate the model parameters for each block, the model tends to be more optimistic when computing the offset. This results in underestimating the influence of lower-priority blocks on the outcome. To tackle this issue, the authors propose to compute the offset using a cross-validation method: first, they split the training set into K equal-size parts $S_1, \cdots, S_K$; then, for individuals in part S_k; they estimated the model parameters using all parts of the training set except S_k and finally, they computed their offsets. The Priority Lasso is described in Algorithm 3.

For this study, we used the Priority-Lasso as implemented in the R package *prioritylasso* (Klau et al., 2020). The chosen block order was RNA, Mutation, and Radiomics.

BlockForest Variable sampling in random forests is biased toward blocks with the largest number of variables. As a result, the small blocks are explored a small relative number of times, regardless of their relationship to the target function. BlockForest adapted the variable sampling to the multi-block structure of data. The various adaptation procedures have already been presented in Chapter 3. We chose to work with the BlockForest sampling method. The procedure starts by choosing a subset of the available J blocks by selecting each block with a probability of 0.5 (this leads to 0.5^J probability

Algorithm 3 Priority-Lasso algorithm

Require: $\boldsymbol{X}^{(1)}, \boldsymbol{X}^{(2)}, \cdots, \boldsymbol{X}^{(J)}, \boldsymbol{Y}, S_1, \cdots, S_K$
$offset(1, \cdots, n) \leftarrow 0$
for $j \in 1, \cdots, J$ **do**
  $model^{(j)} \leftarrow train_model(\boldsymbol{X}^{(j)}, \boldsymbol{Y}, offset)$
  for $k \in 1, \cdots, K$ **do**
    $S_c \leftarrow$ complement of S_k
    $model \leftarrow train_model(\boldsymbol{X}^{(j)}[S_c], \boldsymbol{Y}[S_c], offset)$
    $risk \leftarrow predict_risk(model, X^{(j)}[S_k])$
    $offset[S_k] \leftarrow offset[S_k] + risk$
  end for
end for

that all blocks or no blocks are selected; in the latter case, the sampling is repeated). Then from each chosen block, a fixed number of variables is selected (the number of selected variables is $\sqrt{p^{(j)}}$, where $p^{(j)}$ is the number of variables of the block j). The criterion is then evaluated on the selected variables from each block as in the classical random forest algorithm.

We used the BlockForest as implemented in the R package *blockForest* (Roman Hornung, 2022). We only tested the default parameters except for the parameter splitrule (*splitrule = Log-rank*).

7.1.2 Experimental design

To conduct this study, we used the TCGA-LGG dataset. First, we assess the performance of each block when all data is available. We call this stage the baseline study. This is done to estimate the performance of each block when maximum data is available. Then, we compare the mono-block and multi-block models on only 83 patients having all blocks available. For this study, we used the RNA, Mutation profiles and Radiomics as the three blocks of our study. Figure 7.1 illustrates the described stages.

We used the C-index metric to assess the model performances. The datasets used were split into 80% training set and 20% test set. The training set was used to choose the hyperparameters for each model, including the α and r for the ElasticNet Cox model, the sparsity constraints $s^{(j)}$ and $\lambda_{\mathcal{G}}$ for SGCCA and netSGCCA, and the α for Priority-Lasso. To do so, the training set was split into three equal-sized sets, and a three-fold cross-validation procedure was performed. The chosen parameters are those that yielded the best C-index while selecting less than 10% of variables. The latter condition was added because we are interested in sparse models.

To assess the model performance and evaluate each model variable selection (and its stability), we used ten bootstrapped samples (of equal size to the original training set) with repetition from the training set. Each bootstrapped sample is used to train a model, and its performance is assessed on the isolated test set. The most selected genes are examined from Enrichment Analysis using Enrichr (Chen et al., 2013; Kuleshov et al., 2016; Xie et al., 2021) in order to discover significantly enriched biological pathways. We used the Kyoto Encyclopedia of Genes and Genomes (KEGG), MSigDB hallmarks and MSigDB oncogenic signatures collections. We only considered significantly enriched pathways with a False Discovery Rate (FDR) p-value correction.

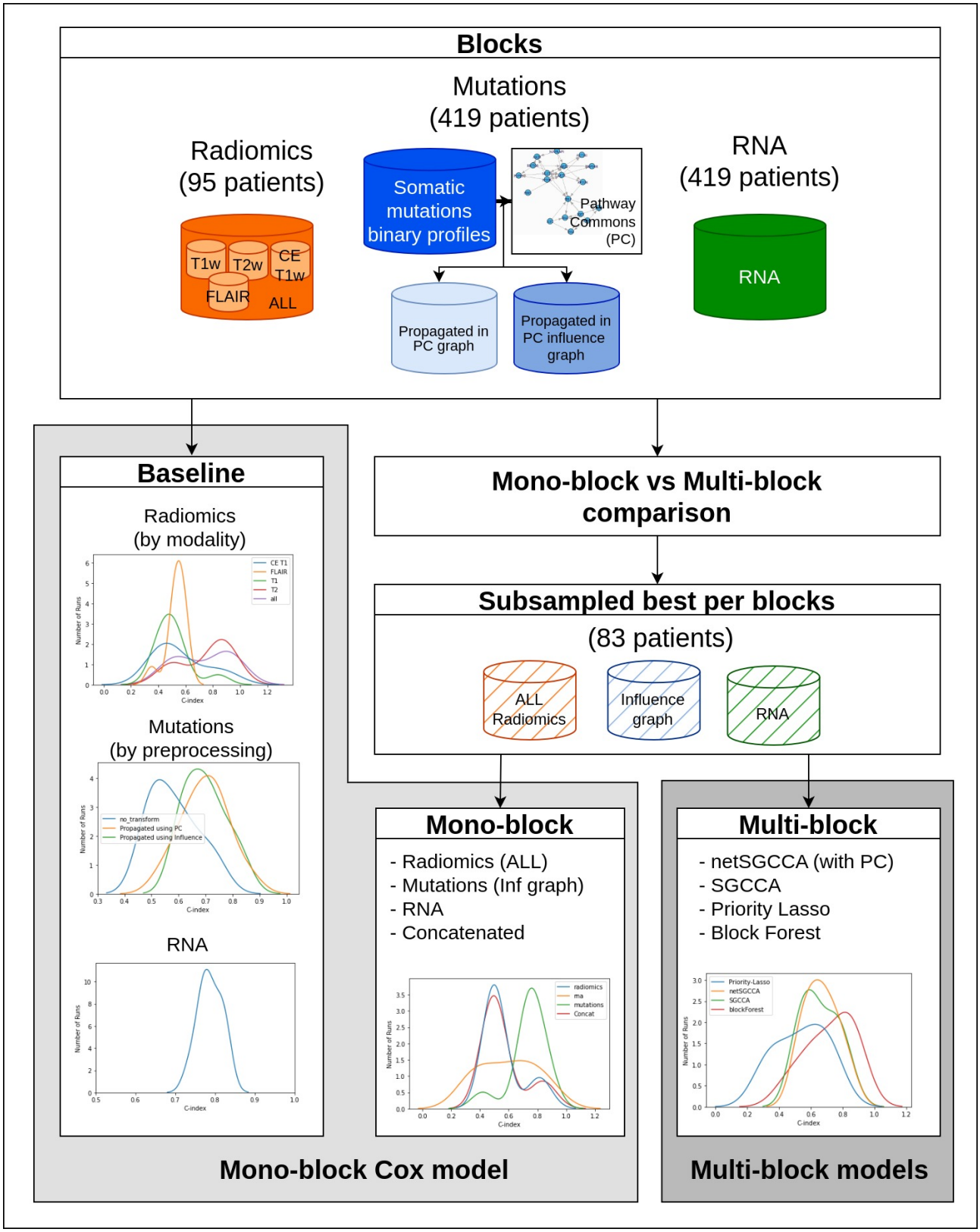


Figure 7.1: Diagram of the different run models and the different comparisons discussed in this chapter.

We have ensured that the overall ratio of censored and non-censored data is kept across all samples and splits. Figure 7.2 illustrates the entire hyperparameter choice and model assessment procedure.

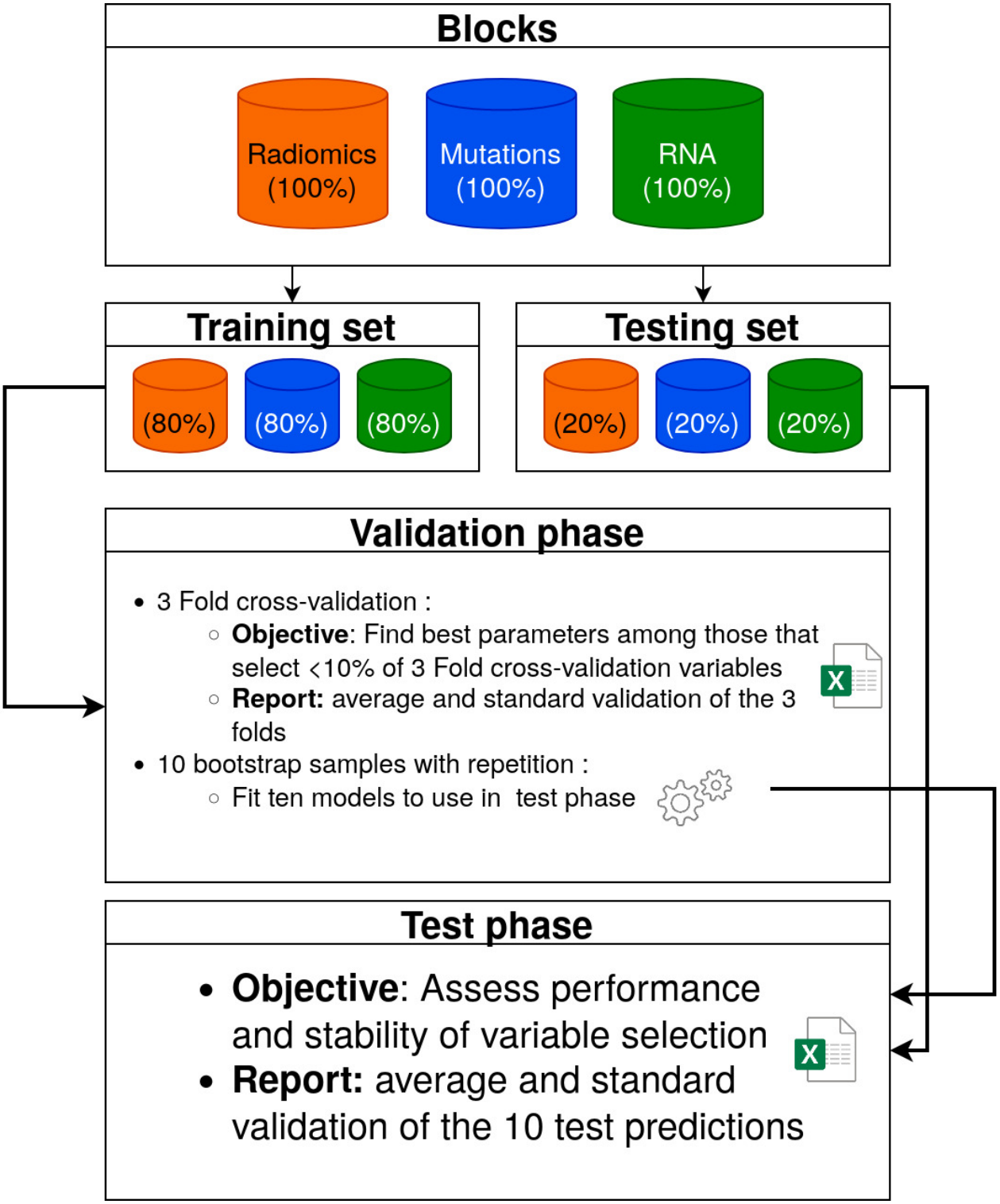


Figure 7.2: Variable selection and model assessment methodology. The same CV folds and bootstrap samples are kept for all compared models.

We used the Cox model implemented in the *glmnet* R package (Friedman et al., 2010; Simon et al., 2011b).

7.2 Baseline results using all per-block available data

The multi-block analysis requires data availability for all subjects across all the blocks. This results in smaller sample sizes when using multi-block approaches compared to mono-block strategies. In order to compare the multi-block and mono-block approaches, we started by studying the predictive performances of each block when all available data is used for this block. This allows us to establish a baseline of the predictive power of each block, from which the multi-block study can be judged. This section does not intend to compare the performances of the different blocks as they are studied across different sets of patients. However, it should show the major differences between the different experimental settings.

7.2.1 Radiomics

Table 7.1 shows the survival prediction results in terms of C-index using the radiomic features on 95 patients from the TCGA-LGG. Figure 7.3 displays the shape of the distributions of the C-index values revealed by the ten bootstrap runs. We discuss further the individual contribution of the radiomics brought by each modality.

Table 7.1: Survival prediction results (C-index) obtained on the 95 patients from the TCGA-LGG dataset using radiomic features. Validation results were obtained on three-fold cross-validation on 76 patients. Test results were obtained using 10 bootstraps on 19 patients. Elastic-Net Cox was used.

	Validation	Test
T1	0.51 ± 0.06	0.51 ± 0.13
T2	0.79 ± 0.06	0.75 ± 0.18
FLAIR	0.69 ± 0.07	0.53 ± 0.07
CE T1	0.64 ± 0.08	0.56 ± 0.19
All	0.80 ± 0.07	0.74 ± 0.20

T1 radiomics. The Cox model could not give good predictions using the T1 modality. On both the validation and test sets, the C-index score remained around 0.5, equivalent to a random ranking of the patients. This result was expected as LGGs are manifested in small areas in the T1-weighted images. The regions of interest (ROI) used to extract the radiomics are dominated by œdema, which is not visible in the T1 weighted images, and provide non-informative radiomics.

T2 radiomics. The Cox model using the radiomics of the T2 weighted modality gave good results on the validation and test sets with an average C-index above 0.75. A closer inspection of the results obtained on the test set shows that 7 of the 10 bootstrap runs scored above 0.84, while the remaining 3 scored 0.5 due to convergence issues. This explains the standard deviation of 0.18 on the test set. The optimal parameters for the T2 radiomics were $\alpha = 0.03$ and $r = 0.85$, which allowed the selection of, on average, 15 variables (except for the three non-convergent samples, where the null model has been returned). Looking only at the seven successful samplings, 64 of the 744 radiomics were selected at least once, and six variables were selected more than three times. Table 7.2 shows the number of times and the average coefficient of these six most selected variables. All features in the table kept the same

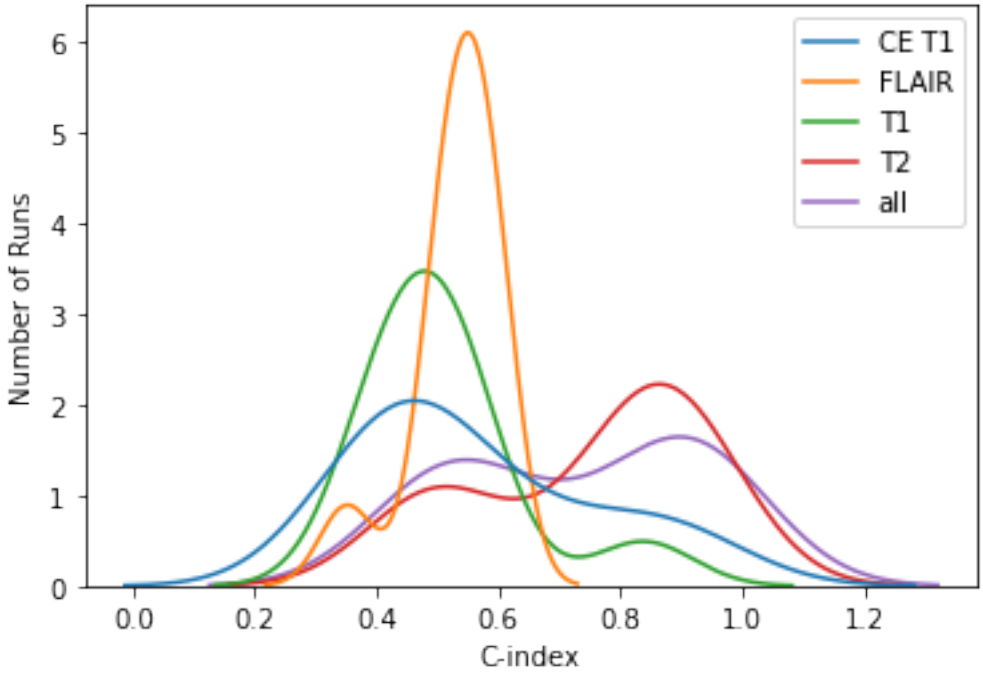


Figure 7.3: Distribution of the C-index values according to the modality of the radiomics. Obtained from the ten bootstrap runs

contribution across all bootstrap models, as indicated by their coefficients not changing signs. Most of these features are extracted from the GLSZM texture matrix, computed on wavelet-transformed images. However, only the "First Order Kurtosis" from the LLL wavelet transformed image has been selected by all models.

Table 7.2: Selected radiomic features extracted from the T2 modality. Out of 10 bootstrap samples, we only kept the seven where the model did not fail. The table only shows radiomic features that were selected more than three times. The means and std were computed using only the samples where the feature was selected.

Selected T2 radiomics	# times selected	coef	min	max
wavelet.LLL_firstorder_Kurtosis_t2	7/7	0.474304 ± 0.162890	0.247739	0.685677
wavelet.HHL_glszm_SmallAreaEmphasis_t2	5/7	-2.678726 ± 2.200181	-5.822927	-0.603556
wavelet.HLH_glszm_SizeZoneNonUniformity.	5/7	-5.288869 ± 4.673160	-10.345156	-1.160420
wavelet.HLH_glszm_GrayLevelNonUniformity_t2	4/7	0.002999 ± 0.002773	0.000355	0.006275
wavelet.LHL_glszm_SmallAreaLowGrayLevelEmphasis_t2	4/7	7.800437 ± 7.900045	1.010237	17.942662
wavelet.LLH_firstorder_Kurtosis_t2	4/7	0.003954 ± 0.002451	0.002106	0.007374

FLAIR radiomics. The radiomics of FLAIR weighted images performed well on the validation set with a C-index of 0.69 but could not generalise, as shown by the results on the test set, where the C-index remained around 0.54. The best-performing parameters were $\alpha = 26.9$ and $r = 0.0039$, meaning the model prioritised regularity over sparsity. On average, each run selected 61 variables, with the lowest selecting only four variables (test C-index of 0.35) and the highest 259 (test C-index of 0.59). The different models selected 316 unique radiomic features (and even 147 when excluding the model which selected 259 variables). Table 7.3 shows the most selected radiomic features. Most selected features are texture features computed from the GLCM, and only the Correlation computed

from the GLCM of the LHL wavelet transformed image was selected in all configurations. The selected variables have lower coefficients compared to the models using the T2 radiomics, which is explained by the optimal low ℓ_1 ratio chosen for the FLAIR radiomics. As with the T2 modality, none of the most selected radiomic features has the sign of their coefficient changed throughout the different runs.

Table 7.3: Selected radiomic features extracted from the FLAIR modality. The table only shows radiomic features that were selected more than 6 times. The means and std were computed using only the samples where the feature was selected.

Selected FLAIR radiomics	# times selected	coeff	min	max
wavelet.LHL_glcm_Correlation_flair	10/10	-0.0124 ± 0.0133	-0.0360	-0.0001
wavelet.LHL_glrlm_RunEntropy_flair	9/10	-0.0047 ± 0.0035	-0.0124	-0.0008
wavelet.LLL_glrlm_GrayLevelNonUniformity	8/10	0.0577 ± 0.0257	0.0253	0.1049
wavelet.LLL_glcm_SumEntropy_flair	8/10	-0.0029 ± 0.0017	-0.0063	-0.0010
wavelet.LLL_glcm_JointEnergy_flair	8/10	0.1065 ± 0.0604	0.0185	0.2133
wavelet.LLL_firstorder_Uniformity_flair	8/10	0.0577 ± 0.0228	0.0282	0.0996
wavelet.LLL_firstorder_Entropy_flair	8/10	-0.0028 ± 0.0017	-0.0064	-0.0010
wavelet.HHL_glcm_Correlation_flair	8/10	-0.0504 ± 0.039	-0.1163	-0.0062
wavelet.LLL_glcm_MaximumProbability_flair	7/10	0.0473 ± 0.0268	0.0158	0.0962
wavelet.LLL_glcm_JointEntropy_flair	7/10	-0.0016 ± 0.001	-0.0034	-0.0004
wavelet.HHL_glcm_InverseVariance_flair	7/10	0.1082 ± 0.099	0.0016	0.2910
wavelet.HLH_glszm_SmallAreaEmphasis_flair	7/10	-0.0202 ± 0.0161	-0.0393	-0.0013
wavelet.HLH_glszm_SizeZoneNonUniformity	7/10	-0.0111 ± 0.0085	-0.0203	-0.0003

CE T1 radiomics. The Cox models trained on the radiomics from the Contrast Enhanced T1 modality (CE T1) have similar results to the FLAIR, with a slightly lower average validation C-index of around 0.64 but a higher test C-index with an average of 0.56. However, as shown in figure 7.3, this test C-index does not reflect a higher ability of the model to generalised compared to the FLAIR, but the instability of the models using the CE T1 radiomic features. Three of the ten samples led to a C-index above 0.7 on the test set, while four scored below 0.43 and thus worst than a random ranking. The models that used CE T1 radiomic selected few variables (about 12), most of them first-order features, and only four were chosen by four or more models, as displayed in table 7.4. The coefficients of the selected variables have small magnitudes, and the models are not stable in their feature selection. Comparing the three runs with the highest C-index on the test set, we observe that selected variables do not overlap except for the "First Order Kurtosis" extracted from the LLL wavelet transformed image. As with the T1, the CE T1 does not show oedema, which can explain why it had difficulties finding robust features.

Table 7.4: Selected radiomic features extracted from the CE T1 modality. The table only shows radiomic features that were selected more than 3 times. The means and standard deviation were computed using only the samples where the feature was selected.

Selected CE T1 radiomics	# times selected	coeff	min	max
wavelet.HHL_firstorder_Skewness_t1Gd	9/10	-0.0077 ± 0.0049	-0.0179	-0.0021
wavelet.HHL_firstorder_Kurtosis_t1Gd	7/10	-0.0001 ± 0.0001	-0.0002	-0.0000
wavelet.LLL_firstorder_Kurtosis_t1Gd	6/10	0.002 ± 0.0004	0.0013	0.0022
wavelet.HHL_glszm_SizeZoneNonUniformity	5/10	-0.0877 ± 0.0698	-0.1787	-0.0036

All modalities radiomics. Finally, we trained Cox models with all four modalities that gave similar results to those when using the T2 modality, with an average C-index of 0.80 for the validation set and 0.74 for the test set. The models trained on radiomic features extracted from all modalities failed to converge twice and returned the null model. Interestingly, only one sample resulted in the model using the T2 and all modalities not converging. This is even though both models were trained in the same sample, which suggests a dependence on the initialisation of the optimisation algorithm. The chosen parameters were $\alpha = 0.085$ and $r = 0.716$, which resulted in 15 variables selected on average. As shown in Table 7.5, the T2 "First Order Kurtosis" extracted from the LLL wavelet transformed image was again the most selected variable, with similar coefficients. This suggests that this radiomic is a good predictor for survival in the case of LGGs. The most selected variables are first-order variables.

Table 7.5: Selected radiomic features extracted from all modalities. Out of 10 bootstrap samples, we only kept the eight with a convergent model. The table only shows radiomic features that were selected more than 3 times. The means and std were computed using only the samples where the feature was selected.

Selected radiomics ALL MODALITIES	# times selected	coeff	min	max
wavelet.LLL_firstorder_Kurtosis_t2	7/8	0.2631 ± 0.1285	0.1080	0.4403
wavelet.LLL_firstorder_Kurtosis_t1Gd	4/8	0.0461 ± 0.0442	0.0050	0.0973
wavelet.HHL_firstorder_Skewness_t1Gd	4/8	-0.1052 ± 0.0996	-0.2531	-0.0379
wavelet.LLL_firstorder_Uniformity_flair	4/8	1.2054 ± 1.1685	0.0655	2.7320

Overall, the T2 and FLAIR modalities performed better than the T1 and CE T1. Using all modalities gave similar performances to those obtained with the T2, but gave different results in terms of variable selection. We hypothesize that the stability issues in the feature selection process are due, on one hand, to the ElasticNet penalty, which can lead to inconsistent results, and on the other hand, to the nature of the radiomic features, which by design contains a lot of redundant information. For the coming comparisons with the multi-block analysis, we used the combination of all modalities because the performance was not degraded while containing more information than the T2 modality alone.

7.2.2 Somatic mutations

We investigated the baseline performance of the Cox model using the mutation matrix. Table 7.6 shows the results obtained when the Cox model was trained on the original or pre-processed mutation profile using the 419 patients from the TCGA-LGG. Figure 7.4 displays the shape of the distributions of the C-index values revealed by the ten bootstrap runs when various pre-processing was applied to the mutation profile. In the following, we compared the results obtained with the raw binary somatic profiles and the profiles that underwent propagation of the mutations according to the Pathway Commons (PC) and its associated influence graph (See 4.2.3).

No transform. The raw mutation profiles performed well in the validation set, with a C-index of 0.73. However, it failed to generalise on the test set, with a C-index of 0.58. The model failed to converge on four runs. This result was expected as the original matrix of mutation profiles is very sparse. Out of the 9235 genes, 5284 are mutated only once, and 9194 are mutated in less than ten patients. Removing the failed runs, we found that the C-index of the remaining six runs is 0.64 with

Table 7.6: Survival prediction results obtained on the 419 patients from the TCGA-LGG dataset using mutation features. Validation results were obtained on three-fold cross-validation. Test results were obtained using 10 bootstraps. Elastic-Net Cox was used

mutation profile pre-processing	Validation	Test
No transform	0.73 ± 0.02	0.58 ± 0.09
Propagated full	0.74 ± 0.01	0.70 ± 0.09
Propagated influence	0.71 ± 0.01	0.70 ± 0.08

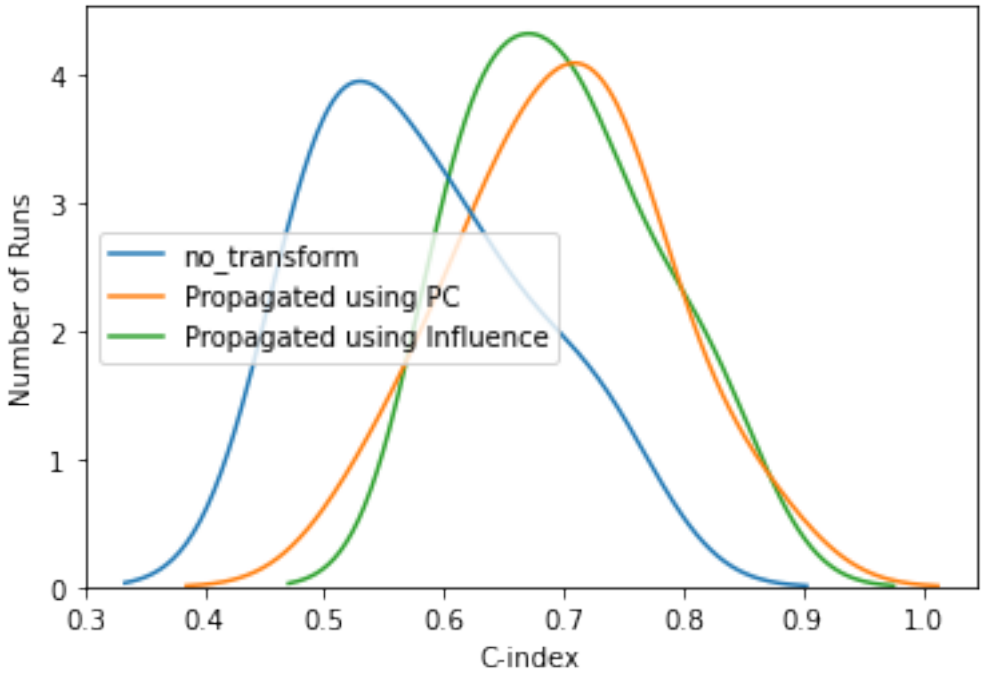


Figure 7.4: Distribution of the C-index values according to the pre-processing of somatic mutation profiles. Obtained from the ten bootstrap runs.

a standard deviation of 0.07. On average, the model selected 876 genes, and 40 genes were selected in all six runs. The list of the 40 selected genes can be found in Table A.1. Enrichment analysis using the Enrichr revealed the oxytocin signalling pathway as the only significantly enriched (with an adjusted p-value of 0.025894) from the Kyoto Encyclopedia of Genes and Genomes (KEGG). This pathway has already been linked to brain tumours and specially LGGs (Liu et al., 2017; Yang and Yang, 2021)

Mutations Propagated according to the PC graph. The models using the mutation matrix after propagation using the Pathway Commons (PC) graph showed similar results to the raw mutation matrix on the validation set, with a C-index of 0.74, but generalised better on the test set with a C-index of 0.70 on average. The chosen parameters were $\alpha = 0.07$ and $r = 0.85$, which resulted in 43 variables selected on average. Using graph propagation helped identify similarities between patients that were not observable beforehand. However, only five genes were selected in more than half the runs. The most selected genes were only mutated in three or fewer subjects, as shown in Figure 7.5. Without propagation, highlighting these genes would have been infeasible. Even though the model identified genes linked to gliomas, no gene was selected consistently, and Enrichr could not find any

significantly enriched pathway.

Table 7.7: Selected genes using the mutation matrix propagated in the PC graph. The table shows genes that were selected more than 5 times. The means and std were computed using only the samples where the feature was selected.

Selected genes - Prop. Full	# times selected	coeff	min	max
PALMD	8/10	144.1463 ± 116.9681	6.0754	296.3065
CHEBI.7815	6/10	-32.797 ± 18.6147	-52.2561	-11.6946
SRSF10	6/10	-0.9108 ± 0.713	-2.0581	-0.1349
MED13	6/10	1.6335 ± 1.1587	0.5644	3.4945
SLCO1A2	6/10	1.7941 ± 1.0129	0.6754	3.3610

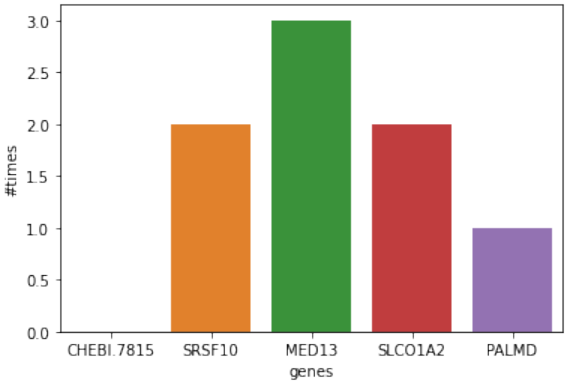
Mutations Propagated according to the PC influence graph. Using the influence graph to propagate the somatic mutations gave similar performances to that obtained with the PC graph, with an average validation C-index of 0.71 and 0.70 for the test set. Additionally, the average number of selected genes is comparable, with 47 genes obtained with the parameters $\alpha = 0.26$ and $r = 0.32$. However, only the gene ***PALMD*** was frequently chosen for both approaches, as shown in Tables 7.8 and 7.7. Interestingly, the gene ***IDH1*** was also selected when using the influence graph. This gene is strongly related to the disease outcome, as discussed earlier. Again, the most frequently chosen genes are only mutated in a small number of patients in the dataset, which shows the importance of graph propagation. Additionally, using the influence graph highlighted multiple pathways, as shown in table 7.9. These pathways involve the genes ***IDH1*** and ***EGFR***, which have already been studied for their link to brain tumours.

Table 7.8: Selected genes using the mutation matrix propagated in the influence graph. The table shows genes that were selected more than 5 times. The means and std were computed using only the samples where the feature was selected.

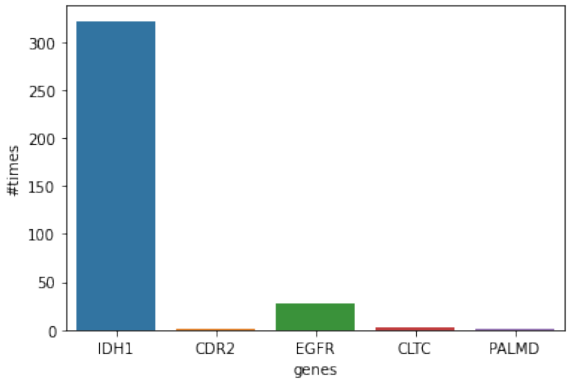
Selected genes - Prop. Influence	# times selected	coeff	min	max
PALMD	10/10	26.5382 ± 18.8677	5.7802	51.9759
IDH1	8/10	-0.2103 ± 0.1712	-0.4126	-0.0006
CLTC	7/10	0.6579 ± 0.4416	0.0949	1.1603
EGFR	6/10	0.3668 ± 0.3568	0.0871	0.8523
CDR2	6/10	-0.1142 ± 0.1035	-0.2337	-0.0033

Table 7.9: Enrichr Enrichment results when using the five most selected genes using the influence graph.

Gene_set	Term	Adjusted P-value	Genes
MSigDB_Hallmark_2020	Protein Secretion	0.001891	CLTC;EGFR
MSigDB_Hallmark_2020	PI3K/AKT/mTOR Signaling	0.001891	CLTC;EGFR
MSigDB_Hallmark_2020	Glycolysis	0.004552	IDH1;EGFR
KEGG_2021_Human	Central carbon metabolism in cancer	0.006356	IDH1;EGFR
KEGG_2021_Human	Endocytosis	0.040868	CLTC;EGFR



PC graph propagation



PC influence graph propagation

Figure 7.5: Number of times the selected genes appear to be mutated in the original somatic mutation profiles, using graph propagation (left) with the PC graph, and (right) with the influence graph.

7.2.3 RNA

We investigated the survival prediction performances of the Cox model when using gene expression data obtained from RNA sequencing obtained on the 419 patients from the TCGA-LGG dataset. Results obtained using RNA data have been the best yet, with a C-index of 0.83 on the validation set and 0.79 in the test set, on average, as shown in Table 7.10 and Figure 7.6. The best-performing parameters were $\alpha = 0.55$ and $r = 0.03$ allowing the selection of 1104 genes on average from 19443 available genes. Among the 11078 uniquely selected genes, only 3079 were selected on over half the runs, and 38 were chosen in all runs (shown in Table A.2). Interestingly, none of these 38 genes coincides with any of the frequently selected genes obtained on the various somatic mutation models. Using the frequently selected genes, we did not find significantly enriched pathways from the KEGG, MSigDB Hallmarks or the MSigDB oncogenic signatures collections.

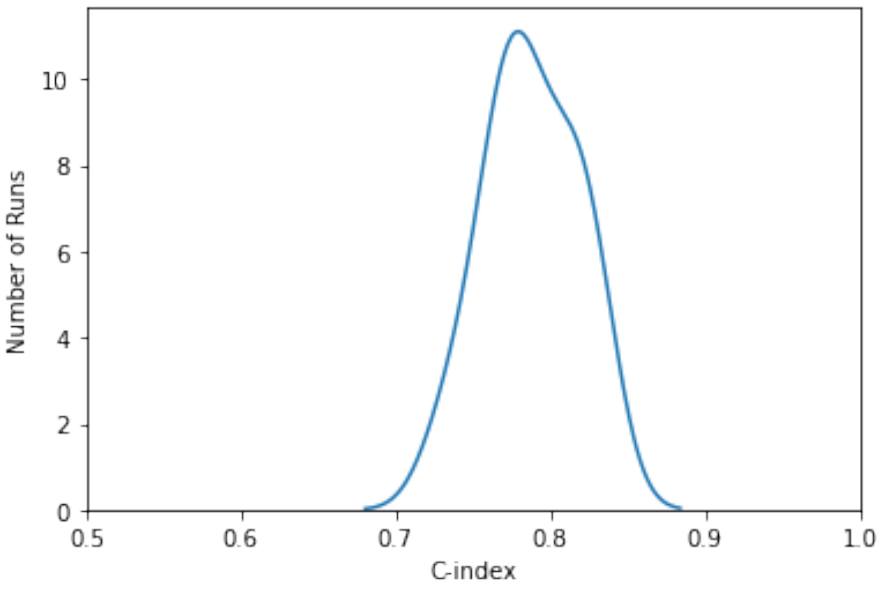


Figure 7.6: Distribution of the C-index values obtained from the ten bootstrap runs for RNA.

Table 7.10: Survival prediction results obtained on the 419 patients from the TCGA-LGG dataset using RNA features. Validation results were obtained on three-fold cross-validation. Test results were obtained using 10 bootstraps. Elastic-Net Cox was used.

	Validation	Test
RNA	0.83 ± 0.02	0.79 ± 0.03

7.3 Comparison mono-block and multi-block approaches

This part is devoted to the comparison of the mono- *vs.* multi- blocks models in terms of performance and interpretation. A reduced sample set is used here consisting of the patients that have all the data available for the multi-block study: 83 subjects have the Radiomics, somatic Mutations and RNA data available in the TCGA-LGG dataset. The different sample sizes are summerized in Table 7.11.

Table 7.11: Sample size comparison between the baseline setting and the current setting devoted to mono-*multi-* block investigation.

	Baseline study (sec. 7.2)	This study (sec. 7.3.1 & 7.3.2)
Radiomics	95	83
Mutations	419	83
RNA	419	83
Blocks concatenated	-	83

7.3.1 Mono-block

We re-ran Cox models using each block (Radiomics, Mutations, RNA) independently to compare the different mono-block with the multi-block approaches. The obtained results from the three folds cross-validation and ten bootstrapping results on the test set are presented in Table 7.12 and Figure 7·7·

Table 7.12: Mono-block results obtained on the 83 subjects having the radiomics, somatic mutations and RNA data available. For the Radiomics, the concatenation of all the radiomics extracted from the four modalities has been used; The somatic mutations with graph propagation with the influence graph was used; Finally, the concatenation of all the blocks was also investigated

	Validation	Test
Radiomics block alone	0.69 ± 0.05	0.56 ± 0.13
Mutations block alone	0.81 ± 0.08	0.74 ± 0.13
RNA block alone	0.75 ± 0.10	0.59 ± 0.20
All blocks Concatenated	0.83 ± 0.14	0.53 ± 0.07

Using the same strategy to tune the models and find the optimal parameters described in the previous section, we compared the radiomics obtained from the four modalities collated together, the somatic mutations profile with the influence graph propagation, the gene expression levels, and finally the concatenation of the three previous blocks.

Radiomics block alone. Using the radiomic features, the Cox model had above random performances on the validation sets with a C-index of 0.69 on average, with a standard deviation of 0.05.

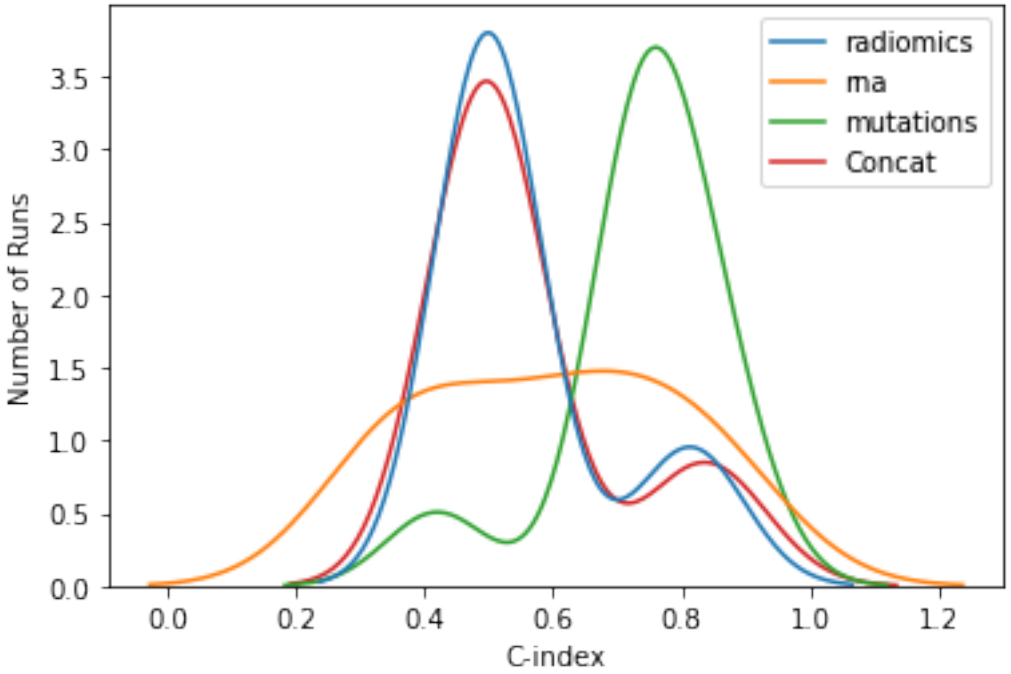


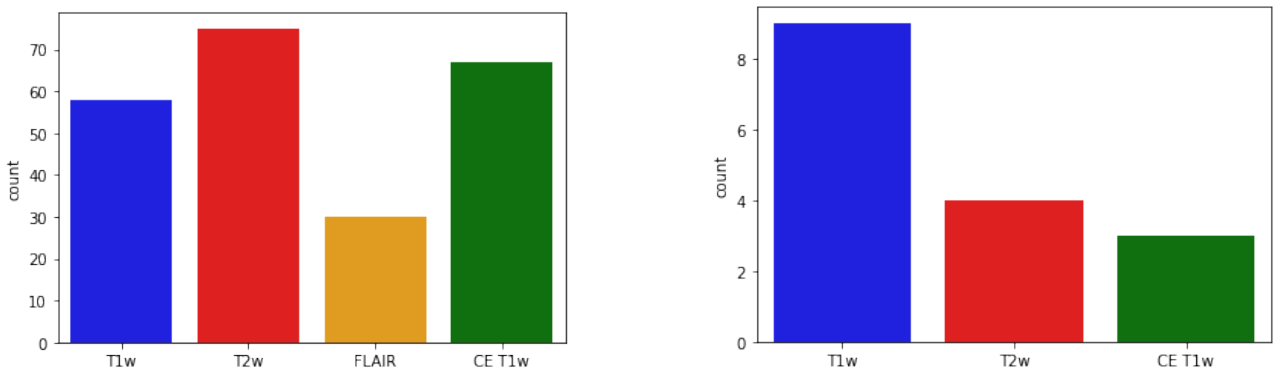
Figure 7.7: Distribution of the C-index values obtained from the ten bootstrap runs for RNA.

However, the model could not generalise on the test set. Seven of the ten bootstrapped runs on the test set failed to converge. On the three remaining runs, two models had a C-index above 0.8, while the remaining one had a C-index of 0.48. The models had an average C-index of 0.71 with a standard deviation of 0.15. The learnt sparsity and regularisation parameters were $\alpha = 0.02$ and $r = 0.25$, respectively. This led to 105 variables selected on average (by taking into account only the three runs which did not return the null model). 230 unique features were selected in at least one of the three runs. Among them, 16 were chosen in the three runs and are described in Table 7.13. Most of the selected covariates are texture features from the GLZM. Most of the consistently selected variables were not identified in the baseline models except for four which have been chosen in the concatenated model. The distribution of the selected radiomic features among the four modalities is shown in Figure 7.8. It displays that the models selected more T2 features in each run. However, the features extracted from the T1 modalities are the most consistently selected, even if the baseline T1 features had poor performances. None of the FLAIR features was consistently selected.

Mutation block alone. Somatic mutation profiles pre-processed with the influence graph showed better results compared to all the other studied data sources. On average, the Cox models had a C-index of 0.81 on the validation and 0.74 on the test sets. The mutation profiles had high and stable results, similar to those obtained with baseline study that used the complete patient sample ($n = 419$ *vs.* $n = 83$). Among the ten bootstrapped runs, nine achieved a C-index above 0.7, while the remaining one only reached a C-index of 0.42. The chosen parameters were $\alpha = 0.72$ and $r = 0.05$, which resulted in 1440 variables selected during each run. The stability of the performances of the Cox models using somatic mutation profiles was not reflected in the stability of chosen variables. 4680 genes were selected in at least one run. Among them, 1225 were selected in more than half of the runs

Table 7.13: Selected radiomic features. The table shows features selected in convergent Cox models. The means and standard deviation were computed using only the samples where the feature was selected.

Selected radiomics	# times selected	coeff	min	max
wavelet.HHL_firstorder_Skewness_t1Gd	3/3	-0.1968 ± 0.0993	-0.2845	-0.0891
wavelet.HHL_firstorder_Kurtosis_t1Gd	3/3	-0.0025 ± 0.0014	-0.0040	-0.0014
wavelet.LLH_glszm_SizeZoneNonUniformityNormalised	3/3	-1.4509 ± 1.025	-2.3291	-0.3246
wavelet.LLL_firstorder_Kurtosis_t2	3/3	0.2224 ± 0.166	0.0307	0.3211
wavelet.HHH_firstorder_RobustMeanAbsoluteDeviation.	3/3	-0.3867 ± 0.4786	-0.9250	-0.0090
wavelet.HHH_firstorder_InterquartileRange_t2	3/3	-0.1844 ± 0.2496	-0.4719	-0.0231
wavelet.HLL_glcm_Imc1_t2	3/3	-8.1397 ± 6.4709	-13.6476	-1.0132
wavelet.HHL_firstorder_Skewness_t1	3/3	-0.0713 ± 0.0672	-0.1462	-0.0161
wavelet.LHL_ngtdm_Strength_t1	3/3	-7.8012 ± 1.3665	-9.3587	-6.8032
wavelet.LHL_glszm_SmallAreaHighGrayLevelEmphasisis	3/3	-0.0006 ± 0.0007	-0.0014	-0.0000
wavelet.LHL_glszm_HighGrayLevelZoneEmphasis_t1	3/3	-0.001 ± 0.0006	-0.0016	-0.0004
wavelet.LHL_glrlm_ShortRunHighGrayLevelEmphasis_t1	3/3	-0.0024 ± 0.0012	-0.0038	-0.0015
wavelet.LHL_glrlm_HighGrayLevelRunEmphasis_t1	3/3	-0.0013 ± 0.0006	-0.0019	-0.0009
wavelet.LHL_gldm_LargeDependenceHighGrayLevelEmphasis	3/3	-0.0 ± 0.0	-0.0000	-0.0000
wavelet.LHL_gldm_HighGrayLevelEmphasis_t1	3/3	-0.0012 ± 0.0005	-0.0019	-0.0009
wavelet.LHL_glcm_Autocorrelation_t1	3/3	-0.0013 ± 0.0006	-0.0019	-0.0009



At least one run All three retained runs

Figure 7.8: Number of radiomic features selected from each modality (left) at least in one run (right) in all three runs.

and only 10 were chosen in all the runs. The ten genes selected in all the runs are shown in Table 7.14. Compared to the baseline, using the 83 individuals resulted in more genes selected. Additionally, all the consistently selected genes were not identified during the baseline phase. Interestingly, three consistently selected genes had coefficients which changed signs between the different runs. This adds inconsistency to the contribution of each gene to the disease outcome. Enrichment analysis using the ten most selected genes revealed four significantly enriched pathways from the KEGG collection, as shown in Table 7.15. These pathways were not identified in the baseline phase but have also been linked to gliomas in previous studies (Guo et al., 2013; Huang et al., 2017; Kim, 2020). These pathways are linked to each other and to the previously identified genes in the baseline.

RNA block alone. Using the gene expression profiles, the Cox model had good performances on the validation set, with a C-index of 0.75 on average. However, the results were lower on the test set across the ten bootstrapped runs. On average, the C-index on the test set was 0.59. As shown in Figure 7.7, three runs had performances below the random results with a C-index around 0.35.

Table 7.14: Selected genes using the somatic mutation profiles. The table shows genes that were selected in all 10 runs. The means and std were computed using only the samples where the feature was selected.

Selected Genes	# times selected	coeff	min	max
SLC27A4_mut	10/10	0.3928 ± 0.1686	0.1346	0.6315
BAAT_mut	10/10	0.2153 ± 0.1183	0.0809	0.4577
SLC13A3_mut	10/10	0.3099 ± 0.2431	0.0130	0.5999
DLG2_mut	10/10	0.0116 ± 0.0528	-0.0205	0.1138
SLC7A9_mut	10/10	0.1258 ± 0.1881	0.0129	0.5459
HIPK4_mut	10/10	0.4212 ± 0.3376	0.0934	1.0119
HSD17B4_mut	10/10	0.0349 ± 0.1029	-0.0214	0.2529
ACOXL_mut	10/10	-0.1394 ± 0.3276	-0.6803	0.1394
ACOT6_mut	10/10	0.3242 ± 0.1227	0.0681	0.4801
DDX41_mut	10/10	0.2847 ± 0.1843	0.0499	0.6012

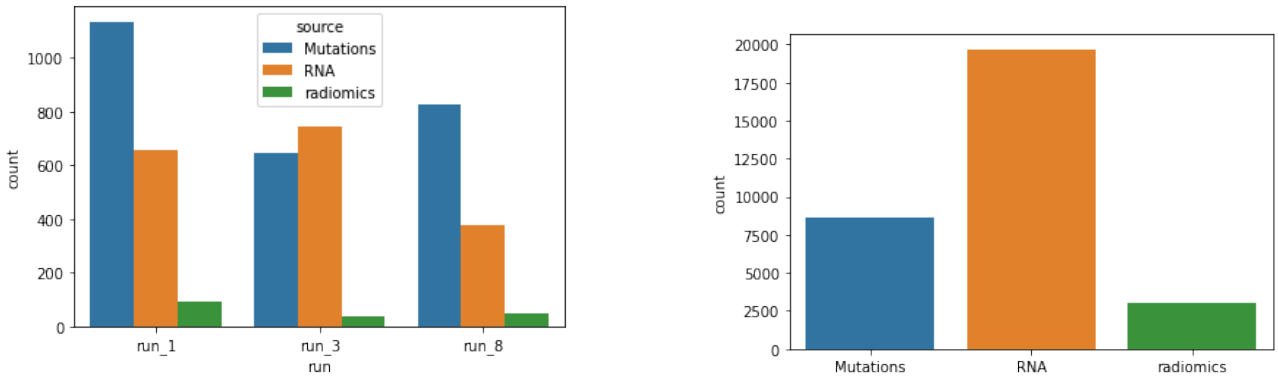
Table 7.15: Enrichment results when using the ten most selected genes using the influence graph.

Gene_set	Term	Adjusted P-value	Genes
KEGG_2021_Human	Primary bile acid biosynthesis	0.000396	HSD17B4;BAAT
KEGG_2021_Human	Biosynthesis of unsaturated fatty acids	0.000510	HSD17B4;BAAT
KEGG_2021_Human	Peroxisome	0.003170	HSD17B4;BAAT
KEGG_2021_Human	Taurine and hypotaurine metabolism	0.017835	BAAT

Meanwhile, five models had a C-index above 0.65. Only one run failed to converge. This exhibits the instability and unreliability of the Cox model when it is fitted with RNA data. The learnt sparsity and regularising parameters were $\alpha = 1.15$ and $r = 0.15$, respectively. This resulted in 13 genes selected on average. However, the variable selection was unstable; most of the genes were selected only once, and only three were selected in more than three runs. Additionally, no gene was selected more than half the runs (five times), which explains the results instability. As a consequence, we do not present any genes from this block.

Radiomics, Mutations and RNA block concatenated. When using all data sources concatenated into a single matrix, the Cox model performed well on the validation set, with a C-index of 0.83. However, as seen by the bootstrapped test runs, the models failed to generalise, which obtained an average C-index of 0.53. Among the ten runs, only three did not result in the null model, and their C-indices were 0.81, 0.47 and 0.86. The learnt sparsity and regularizing parameters were $\alpha = 2.53$ and $r = 0.01$, respectively. This resulted in 1521 selected variables on average (only considering the three successful runs). As expected, most selected variables came from the Mutation block, which has shown the best result so far, as shown in Figure 7.9(a). We note that run_3 had the lowest C-index, which corresponds to the run where the Cox model using the mutations alone would not generalise. Among all the variables selected, 262 were present on the three successful runs. Enrichment analysis on consistently selected variables from the mutation block highlighted the Phototransduction and Purine metabolism pathways.

Summary. In general, removing subjects from the datasets considerably changes the performance of trained model especially as regards its generalization capabilities and selection stability. RNA



((a)) ((b))

Figure 7.9: (a) Number of selected variables from each block, in each successful run using the Cox model with all blocks concatenated. (b) Total number of variables from each block using the Cox model with all blocks concatenated

Table 7.16: Enrichment results when using the matrix with concatenated variables from all data sources.

Gene_set	Term	Adjusted P-value	Genes
KEGG_2021_Human	Phototransduction	0.019943	PDE6G;PDE6B;PDE6A;GUCY2F
KEGG_2021_Human	Purine metabolism	0.019943	GUCY2C;NPR1;PDE6G;PDE6C; PDE6B;PDE6A;GUCY2F

data was the best-performing block in the baseline but performed poorly when only 83 subjects were kept from the whole dataset. Furthermore, mutation results gave similar performances between the baseline and the reduced mono-block model. However, the selected genes and their significant enriched pathways differed. For radiomic features, the results from the baseline and this mono-block study were similar, which was expected given the low number of removed subjects. Additionally, we remark that the concatenated results reflected the results obtained on the individual blocks.

The mono-block Cox model was generally unstable in terms of variable selection. Additionally, multiple runs failed to converge even with similar samples and parameters. The Cox model appears to be initialisation dependent, which weakens its reliability.

As an interim conclusion, these results question the reliability of the mono-block Cox model we obtained. It also raises questions about the interpretations drawn from this model that learns in a mere (concatenation) mono-block with the classical ElasticNet penalty.

7.3.2 Multi-block

In this section, we compared the different multi-block models using the 83 subjects having all data blocks available. The first two multi-block models we considered here are SGCCA, netSGCCA; They are extensions of the RGCCA framework that include respectively sparsity (Tenenhaus et al., 2014) and graphical constraint developed in section 6.3. We compared these two models to two other approaches established independently: Priority-Lasso and BlockForest(see 7.1.1). The cross-validation and bootstrapped sets are the same as those used in the mono-block study. This is done to obtain comparable results. Similarly to the mono-block study, we used the RNA block, the radiomic features obtained from the four modalities and the somatic mutation profile processed with the influence graph

propagation. Results are summed up in Table 7.17. The C-index scores distribution on the test set is presented in Figure 7.10.

Table 7.17: Multi-block results obtained on 83 subjects having the radiomics, somatic and RNA data available. For the radiomics, the concatenation of all the radiomics extracted from the four modalities has been used; The somatic mutations with graph propagation with the influence graph was used.

	Validation	Test
netSGCCA	0.88 ± 0.14	0.67 ± 0.10
SGCCA	0.88 ± 0.14	0.66 ± 0.12
Priority-Lasso	0.71 ± 0.11	0.54 ± 0.16
BlockForest	0.68 ± 0.06	0.72 ± 0.15

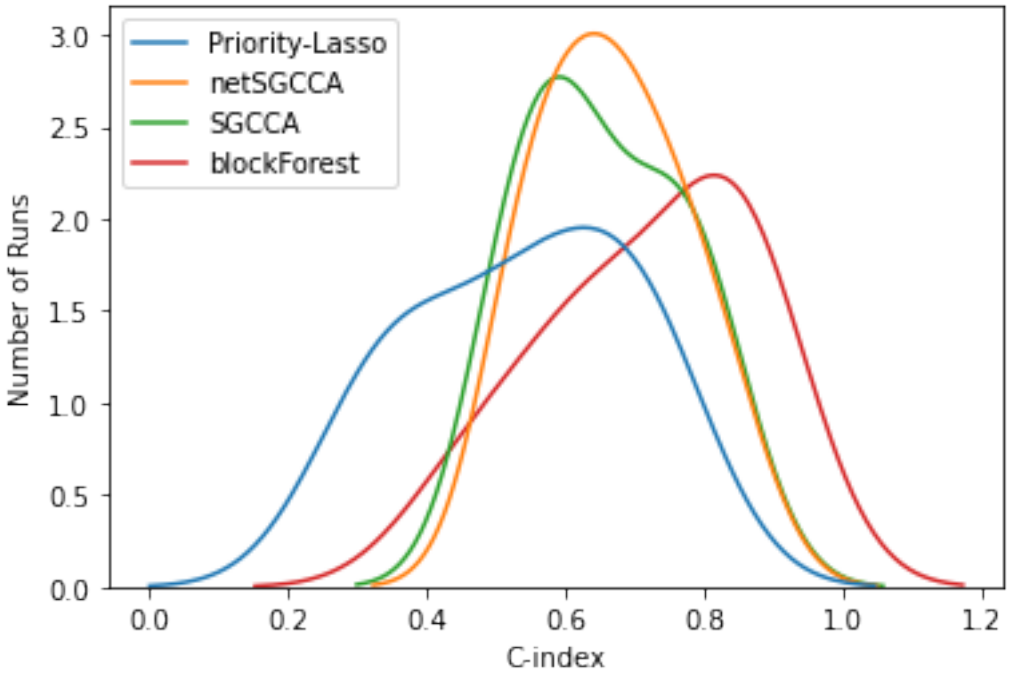


Figure 7.10: Distribution of the C-index values obtained from the ten bootstrap runs, for the four multiblock models tested.

SGCCA and netSGCCA. Using the netSGCCA with the normalised Pathway Common graph penalty on the RNA, we obtained high performances on the validation set, with an average C-index of 0.88. However, the performance dropped to 0.67 for the test set, which is higher than most mono-block models but below the mono-block model based on the Mutations. On average, in each test run, the model selected 1800 variables. Among them, around 47 are selected from the Mutations block, 1752 from the RNA block and one variable from the radiomics.

Table 7.18 shows the 17 most frequently selected features from the RNA block. It exhibits the instability of the model in terms of variable selection. Two reasons can explain this. First, as shown earlier, the RNA block is not a good predictor for survival in this setting sample ($n = 83$, see also 7.3.1): Even if the GraphNet penalty allows more variables to be selected from this block, we can not expect the RNA block to yield stable selected variables. Second, the GraphNet penalty groups

correlated features and using the bootstrap changes the correlation profiles between the genes, as shown in Figure 7.11. The correlation profile of the gene ***PGPEPE1*** for example, is volatile across the folds. The instability of variable selection on the other blocks is somewhat inherited from the instability of the SGCCA, which has already been highlighted in the previous chapter.

Enrichment analysis results of the 17 most selected genes (RNA block) are reported in table 7.19. The ***IL15*** gene set has immune functions and has already been shown to improve anti-glioma activity (Krenciute et al., 2017). The KRAS has been linked to increased risk of LGGs (Guan et al., 2021; Ryall et al., 2020). PKCA has long been identified as a key regulator of the growth of malignant gliomas (Arcos-Montoya et al., 2021; Baltuch et al., 1995; Cameron et al., 2008).

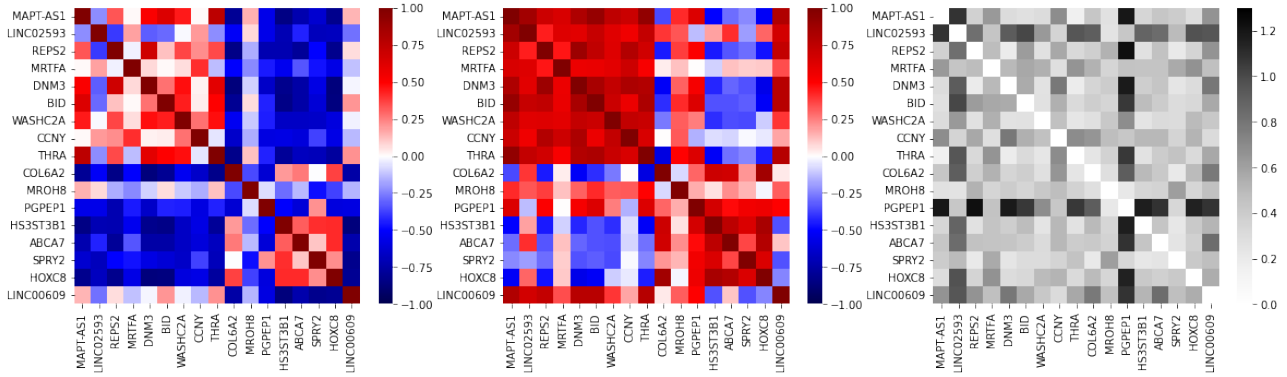


Figure 7.11: Correlation matrices between the 17 most frequently selected genes from the RNA block by the netSGCCA model. (left panel): the minimum correlation between the genes in the ten bootstrap samples; (center panel): the maximum correlation between the genes in the ten bootstrap samples; (right panel): the difference between the two previous matrices.

Table 7.18: Selected features using the netSGCCA, the normalised Pathway Commons Laplacian was applied on RNA data. The table shows features that were selected more than 5 times. The means and std were computed using only the samples where the feature was selected.

Selected radiomics	# times selected	coeff	min	max
THRA	6/10	0.0001 ± 0.0015	-0.0023	0.0018
BID	5/10	-0.0009 ± 0.0023	-0.0037	0.0026
LINC02593	5/10	-0.0236 ± 0.0285	-0.0480	0.0180
DNM3	5/10	-0.0011 ± 0.0027	-0.0054	0.0017
MRTFA	5/10	-0.0012 ± 0.0019	-0.0041	0.0012
REPS2	5/10	-0.0017 ± 0.0032	-0.0074	0.0006
COL6A2	5/10	0.0004 ± 0.0023	-0.0029	0.0033
SPRY2	5/10	0.0018 ± 0.0053	-0.0050	0.0097
LINC00609	5/10	0.0027 ± 0.0428	-0.0402	0.0566
PGPEP1	5/10	0.0007 ± 0.0008	0.0001	0.0021
HS3ST3B1	5/10	0.0008 ± 0.0015	-0.0010	0.0030
CCNY	5/10	-0.0008 ± 0.0009	-0.0024	0.0001
MROH8	5/10	0.0005 ± 0.0019	-0.0012	0.0032
WASHC2A	5/10	-0.0009 ± 0.0019	-0.0028	0.0020
ABCA7	5/10	0.0011 ± 0.0023	-0.0010	0.0047
HOXC8	5/10	0.0024 ± 0.0041	-0.0027	0.0086
MAPT-AS1	5/10	-0.0249 ± 0.0462	-0.0733	0.0499

Table 7.19: Enrichment results when using the most selected genes from the RNA block using netSGCCA.

Gene_set	Term	Adjusted P-value	Genes
MSigDB_Oncogenic_Signatures	IL15 UP.V1 UP	0.019308	HS3ST3B1;SPRY2;HOXC8
MSigDB_Oncogenic_Signatures	KRAS.600 UP.V1 UP	0.030833	HS3ST3B1;DNM3;SPRY2
MSigDB_Oncogenic_Signatures	KRAS.AMP.LUNG UP.V1 DN	0.047247	PGPEP1;HOXC8
MSigDB_Oncogenic_Signatures	KRAS.300 UP.V1 UP	0.047247	DNM3;SPRY2
MSigDB_Oncogenic_Signatures	CSR EARLY UP.V1 UP	0.047247	REPS2;SPRY2
MSigDB_Oncogenic_Signatures	PKCA DN.V1 DN	0.047247	BID;HOXC8
MSigDB_Oncogenic_Signatures	AKT UP.V1 UP	0.047247	ABCA7;BID

Further analysis revealed that removing the radiomic block from the multi-block netSGCCA model increased the C-index on the test runs, with an average score of 0.8. It should be noted that we did not run proper hyperparameters selection for these two models, thus the results are only indicative and not included in the result tables. On the other hand, using only the RNA and radiomic blocks yielded a lower average C-index of 0.62 on the test set.

SGCCA coupled with a Cox model had a similar performance as the one obtained with netSGCCA coupled with a Cox model on the validation and test sets, with a C-index of 0.88 and 0.66, respectively. However, the model with SGCCA was more sparse on the RNA block by selecting, on average, only three variables. Variable selection on the other blocks did not meaningfully change between SGCCA and netSGCCA. These results confirm previous conclusions discussed in the last chapter.

Priority-Lasso and BlackForest. Using Priority-Lasso, we obtained an average C-index of 0.71 on the validation set. After a closer inspection of the obtained performances on the validation set, we observed an average C-index of 0.70 with a standard deviation of 0.14 using the RNA alone. Adding the mutations did not significantly change the results, with an average C-index of 0.71 and a standard deviation of 0.15. Adding the radiomic features did not change the average C-index from 0.71; however, the standard deviation decreased to 0.11. Radiomic features helped the stability of the model. Priority-Lasso could not generalise the results obtained on the validation set into the test set with a mean C-index of 0.54 on the ten bootstrapped runs. Among the ten runs, three performed below 0.4, which is worst than average. The instability is reflected in the selected variables. In the ten runs, 50 unique variables were selected (with an average of 7.6 variables per run), and no variable was selected in more than six runs. Priority-Lasso only uses a Lasso penalty which does not add regularity to the model estimates, which explains the low number of selected variables and their instability.

Compared to the other multi-block models, BlockForest performed the best on the test set, with a C-index of 0.68 on the validation set and 0.72 on the test set. netSGCCa and SGCCA gave slightly lower but comparable results. However, Block Forest does not allow variable selection, which hinders the interpretability of the model and the extraction of genes of interest.

7.4 Discusion

The netSGCCA was among the best-performing models on the validation and test set. The C-index was stable during all test runs, which was not the case for the mono-block Cox models (see 7.3.1). The netSGCCA model was only topped by BlockForest among all multi-block models. However, it should be noted that the BlockForest is an unconstrained model and does not perform variable selection,

which is a key part of this study. The netSGCCA model was also able to select genes and pathways already linked to the development of gliomas and the prognosis of LGGs.

The radiomic integration, using the netSGCCA, confirmed the results obtained in the previous chapter. The graph penalty allows the selection of more variables, and the correlation is the main driver of the grouping effect. Furthermore, the netSGCCA and SGCCA had similar performances when the Pathway Commons (PC) graph was used to construct the penalty. This was also a previously obtained result. Yet, the baseline results demonstrate that radiomic features can be good predictors of the survival of Lower Grade Glioma. The T2w modality had especially high performances. This has already been observed in the literature. Unfortunately, we could not recover obtained results in the baseline when using the radiomics on the 83 patients having all blocks available. The First Order Kurtosis from the T2w modality was the best predictor among all radiomic features. A similar result has already been established in the chapter of related works Liu et al. (2018), though the feature was computed on a different wavelet-transformed image.

Overall, data availability restriction hindered the predictive performance of the blocks. Furthermore, the Cox model became less stable when data from only 83 patients were used. Additionally, we expected an overlap between the variables selected on the baseline and on the restricted dataset, however results obtained show otherwise. The netSGCCA showed promising results in term of survival prediction and variable selection.Therefore, further studies are required to validate the results. It is also necessary to perform this study on a larger dataset.

7.5 Conclusion

The netSGCCA showed to be a robust model able to select variables already linked to the studied pathology and interacting in relevant biological pathways. Adding the imaging features hindered the ability of the netSGCCA to extract features that can be used for survival prediction. However, the baseline results show that radiomic features extracted from the T2w modality can be strong predictors. Thus, larger datasets are required to further investigate the potential of radiomic features. The usage of graphs was shown to be impactful both when used to propagate somatic mutations and as a penalty in the netSGCCA. Our results show that multi-block integration under the netSGCCA procedure can uncover candidate genes and biomarkers, which can be investigated in later studies. Our framework could be used to study the DIPG and discover new relevant biomarkers.

* * *
* *
*

Chapter 8

Conclusion

Our work was motivated by the study of the Pediatric Diffuse High Grade Gliomas (DIPG), a rare pediatric high grade glioma. We aimed to design statistical methods able to integrate imaging and genetic data. These methods must be interpretable, thus giving new insights into the genetic signature of the tumour and its link to imaging features.

Radiomic analysis requires predefined Regions of Interest (ROI) on the images at hand. For our DIPG cohort, manual segmentation of the tumour was not feasible. Additionally, no database has been built to train classical machine-learning algorithms to delineate the tumoural regions automatically. This study focused on obtaining binary segmentations for the DIPG using only the FLAIR and T2w modalities, using models trained on Glioblastoma. We propose combining different simple models of detection and segmentation to obtain satisfying segmentation results on a cohort that contained differences compared to the training dataset regarding, among others, patient age, image quality and tumour type.

In parallel, our work aimed at building a multi-block integration model which takes into account known complex graphs of interactions between the genes. Additionally, our focus was on understanding how the graph of interactions influences variable selection. We propose netSGCCA, a model combining the Sparse Generalised Canonical Correlation Analysis (SGCCA) and the GraphNet penalty. We applied our model to the publicly available TCGA-LGG dataset. Contrary to Elastic-Net alone, GraphNet penalty can select a reasonable set of genes and yields informative biological interpretation from the pathway enrichment analysis. The example on the TCGA-LGG dataset exhibits the stability and reliability of netSGCCA for selecting variables of interest. However, it is essential to note that we show that the co-selection of variables is not primarily influenced by the structure of the graph but rather by its overall density. Therefore, an interpretation in terms of the pathways selected from the graph is elusive. Nevertheless, the method did extract genes that have been found co-(de)regulated in other studies of low-grade gliomas and other brain tumours.

Finally, we used netSGCCA to integrate radiomics and genetic data and applied it to the task of survival prediction. Due to the unavailability of survival data on our DIPG cohort, we used the TCGA-LGG dataset to conduct the study. We compared the results obtained with netSGCCA with other multi-block survival approaches and the mono-block built models. Our study focused on the variables selected by the model and compared the results with the available literature. The netSGCCA showed to be a robust model able to select variables already linked to the studied pathology and interacting

in relevant biological pathways. Adding the imaging features hindered the ability of the netSGCCA to extract features that can be used for survival prediction. However, the baseline results show that radiomic features extracted from the T2w modality can be strong predictors. Thus, larger datasets are required to further investigate the potential of radiomic features.

Future works

Studying rare tumours with statistical models is challenging. Data is scarce due to the nature of the tumour. Larger datasets are required to obtain conclusive and robust results. We also observe a high variability in quality and a high number of missing data, which is due to the fact that the data are collected in a clinical setting. The difficulty of adhering to a strict protocol designed for statistical analysis is obviously understood. It is important to carefully handle these datasets, as the available processing tools and models are not robust to noisy, high-variability data.

Our work proposed a segmentation approach that allowed us to obtain ROIs on the DIPG. However, further work should focus on finely delineating the tumours and their sub-components. Furthermore, the quality of segmentation and its impact on the radiomic features must be assessed in detail.

Missing data is inherent to multi-block studies, as collecting data from multiple sources in a clinical context is definitely hard to achieve. This thesis did not explore handling missing data in multi-block integration, which resulted in using a relatively small dataset and would deserve future work.

Finally, this work proposed using radiomic features in a multi-block framework. However, these features were biased towards the acquisition tools, reconstruction algorithm and preprocessing methods. No definitive conclusion can be drawn about image signatures of the tumours without strict control of all variability sources. Future works should exploit, for example, the already available follow-up MRI data when the embargo on the clinical annotations of the clinical study is raised.

* * *
* *
*

Appendix A

First Appendix

Table A.1: Selected mutated genes using the raw mutation matrix. Out of 10 bootstrap samples, we only kept the 6 where the model did not fail. Table only shows genes that were selected on all the runs.

Selected genes	# times selected	coeff	min	max
CLTC	6/6	0.4685 ± 0.1167	0.2736	0.5710
EDA	6/6	-0.0528 ± 0.0107	-0.0664	-0.0404
NLRP9	6/6	0.2261 ± 0.1934	0.0211	0.5266
SLC45A4	6/6	0.1353 ± 0.1356	0.0348	0.3950
NAA11	6/6	0.1733 ± 0.1833	0.0221	0.4143
PHRF1	6/6	-0.1101 ± 0.1325	-0.3799	-0.0435
KCNS3	6/6	0.195 ± 0.1929	0.0229	0.4788
PNMA3	6/6	0.167 ± 0.1359	0.0020	0.3889
ZBED9	6/6	0.2607 ± 0.114	0.1019	0.3695
NEBL	6/6	0.3167 ± 0.171	0.0526	0.5574
SLC11A1	6/6	0.4093 ± 0.2297	0.0117	0.5964
EDAR	6/6	0.5349 ± 0.0856	0.4519	0.6780
SRSF10	6/6	-0.3321 ± 0.2118	-0.6519	-0.1169
AFP	6/6	0.1447 ± 0.1228	0.0000	0.3120
MED13	6/6	0.3441 ± 0.1756	0.1252	0.5231
IGHG4	6/6	0.2067 ± 0.1587	0.0524	0.3725
OR1N2	6/6	0.0988 ± 0.1107	0.0202	0.3194
EGFR	6/6	0.2044 ± 0.1664	0.0248	0.4153
FLG	6/6	0.2363 ± 0.1303	0.0959	0.4043
CLMN	6/6	-0.2639 ± 0.1702	-0.5105	-0.1151
SLC4A3	6/6	0.2097 ± 0.1922	0.0091	0.4483
ELK1	6/6	0.1748 ± 0.1876	0.0250	0.4886
SYBU	6/6	-0.0536 ± 0.0107	-0.0669	-0.0405
PDCD2L	6/6	-0.0529 ± 0.0107	-0.0665	-0.0403
IGHV3.43	6/6	-0.053 ± 0.0106	-0.0666	-0.0403
RP11.159L20.2	6/6	-0.0531 ± 0.0106	-0.0666	-0.0402
OR5B21	6/6	-0.0534 ± 0.0105	-0.0667	-0.0401
OR52D1	6/6	-0.0534 ± 0.0105	-0.0667	-0.0402
TBATA	6/6	-0.0535 ± 0.0106	-0.0668	-0.0402
MASTL	6/6	-0.0535 ± 0.0106	-0.0669	-0.0404
MIR133B	6/6	-0.0536 ± 0.0108	-0.0670	-0.0407
GPBP1L1	6/6	0.2042 ± 0.1514	0.0659	0.4142
PRKAA1	6/6	-0.0536 ± 0.0109	-0.0670	-0.0409
ORC4	6/6	-0.0535 ± 0.0111	-0.0671	-0.0411
ACBD3	6/6	-0.0536 ± 0.0112	-0.0671	-0.0413
ALPK3	6/6	0.2049 ± 0.1899	0.0403	0.4929
PCDHGA5	6/6	0.2339 ± 0.1979	0.0396	0.5050
EPN3	6/6	0.2596 ± 0.1542	0.0450	0.4117
PPP3R2	6/6	0.2117 ± 0.1243	0.0161	0.3088
IDH1	6/6	-0.1428 ± 0.0543	-0.2377	-0.0916

Table A.2: Selected genes using the RNA data. Table shows genes that were selected in all runs.

Selected genes	# times selected	coeff	min	max
SLC28A2.AS1	10/10	-0.0155 ± 0.0071	-0.0259	-0.0071
PLAA	10/10	-0.0553 ± 0.0203	-0.0880	-0.0258
DZIP1L	10/10	0.0202 ± 0.0125	0.0041	0.0384
BTG3	10/10	0.0102 ± 0.0056	0.0021	0.0208
SFRP2	10/10	-0.0047 ± 0.0022	-0.0091	-0.0027
COL8A1	10/10	0.0055 ± 0.0022	0.0016	0.0100
ENTPD1	10/10	-0.0341 ± 0.0203	-0.0755	-0.0058
SLC28A2	10/10	-0.0271 ± 0.0078	-0.0366	-0.0125
FHDC1	10/10	-0.0089 ± 0.0047	-0.0157	-0.0022
WDR38	10/10	0.0035 ± 0.0017	0.0005	0.0057
RHOQ.AS1	10/10	0.0084 ± 0.0071	0.0020	0.0235
LTV1	10/10	-0.0301 ± 0.0155	-0.0487	-0.0029
CASP9	10/10	-0.0261 ± 0.0058	-0.0341	-0.0182
ELL3	10/10	0.0154 ± 0.0097	0.0014	0.0355
H2BC11	10/10	0.0084 ± 0.0045	0.0018	0.0168
CAAP1	10/10	-0.0355 ± 0.0173	-0.0623	-0.0157
NEK6	10/10	0.0093 ± 0.0067	0.0006	0.0187
MTAP	10/10	-0.0239 ± 0.0106	-0.0391	-0.0081
NUAK2	10/10	0.0182 ± 0.0095	0.0029	0.0359
SLC9B1	10/10	-0.0303 ± 0.0158	-0.0661	-0.0100
TNFRSF11B	10/10	0.0062 ± 0.0036	0.0016	0.0132
CCDC178	10/10	-0.0166 ± 0.0073	-0.0300	-0.0100
LINC02198	10/10	0.0113 ± 0.005	0.0045	0.0188
LINC02044	10/10	-0.0138 ± 0.0101	-0.0356	-0.0013
FLG.AS1	10/10	-0.0081 ± 0.0039	-0.0157	-0.0040
HLA.DQB2	10/10	0.006 ± 0.0027	0.0012	0.0092
EHBP1.AS1	10/10	0.0116 ± 0.0089	0.0010	0.0280
CBWD4P	10/10	-0.027 ± 0.011	-0.0436	-0.0107
MARCHF5	10/10	-0.0362 ± 0.0141	-0.0514	-0.0099
LINC00994	10/10	-0.0092 ± 0.0034	-0.0137	-0.0028
H1.9P	10/10	0.0064 ± 0.0018	0.0034	0.0098
NUTM2A	10/10	-0.0141 ± 0.0106	-0.0321	-0.0004
ZNF648	10/10	0.01 ± 0.0049	0.0006	0.0168
GOLGA6L2	10/10	-0.0075 ± 0.0038	-0.0129	-0.0005
TDRD12	10/10	-0.0149 ± 0.0084	-0.0316	-0.0046
MLLT3	10/10	-0.0424 ± 0.0179	-0.0622	-0.0114
PLEKHA2	10/10	0.0309 ± 0.0177	0.0090	0.0652
KCNN2	10/10	-0.0106 ± 0.0076	-0.0232	-0.0002

Appendix B

Second Appendix

We want to prove the grouping effect of GraphNet in the context of netSGCCA. Our proof is similar to Zou and Hastie (2005) and Li and Li (2008). We assume that two variables u and v are only connected to each other in the *a priori* graph. We also assume that u and v have similar contributions (i.e. their respective weights have the same sign). Give the lagrangian equation defined in 6.6:

$$\begin{aligned} &\underset{\boldsymbol{w}^{(1)},\boldsymbol{w}^{(2)},\cdots,\boldsymbol{w}^{(J)}}{\operatorname{argmin}} h(\boldsymbol{w}^{(1)},\boldsymbol{w}^{(1)},\cdots,\boldsymbol{w}^{(J)}) \\ h(\boldsymbol{w}^{(1)},\boldsymbol{w}^{(2)},\cdots,\boldsymbol{w}^{(J)}) &= \sum_{\substack{j,k=1\\ j\neq k}}^{J} -c_{j,k}\operatorname{cov}\left(\boldsymbol{X}^{(j)}\boldsymbol{w}^{(j)},\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)}\right) + \frac{\gamma_{\mathcal{G}}}{\lambda_{\max}}\boldsymbol{w}^{(g)\top}\boldsymbol{L}_{\mathcal{G}}\boldsymbol{w}^{(g)} \\ &\quad + \sum_{j=1}^{J}\lambda_1^{(j)}\|\boldsymbol{w}^{(j)}\|_1 + \sum_{j=1}^{J}\lambda_2^{(j)}\|\boldsymbol{w}^{(j)}\|_2^2 \end{aligned} \tag{B.1}$$

Under the optimality conditions of $\boldsymbol{w}^{(g)}$ (the weight of the block on which GrahNet is applied), we have:

$$\begin{aligned} &\frac{\partial h}{\partial \boldsymbol{w}^{(g)}} = 0 \\ \text{Which implies} \quad &-\boldsymbol{X}^{(g)\top}Z + \frac{\gamma_{\mathcal{G}}}{\lambda_{\max}}\boldsymbol{L}_{\mathcal{G}}\boldsymbol{w}^{(g)} + \lambda_1^{(g)}\operatorname{sign}(\boldsymbol{w}^{(g)}) + 2\lambda_2^{(g)}\boldsymbol{w}^{(g)} = 0 \\ &\text{s.t.} \quad Z = \sum_{\substack{k=1\\ j\neq k}}^{J} c_{g,k}\boldsymbol{X}^{(k)}\boldsymbol{w}^{(k)} \end{aligned} \tag{B.2}$$

For the variables u and v, let $\boldsymbol{x}_u$ and $\boldsymbol{x}_v$ be the columns v and u from the matrix $\boldsymbol{X}^{(g)}$. For these two variables, the equality becomes :

$$\begin{aligned} \text{For } u: \quad &-\boldsymbol{x}_u^\top Z + \frac{\gamma_{\mathcal{G}}}{\lambda_{\max}}\boldsymbol{L}_{\mathcal{G}}(\boldsymbol{w}_u - \boldsymbol{w}_v) + \lambda_1^{(g)}\operatorname{sign}(\boldsymbol{w}_u) + 2\lambda_2^{(g)}\boldsymbol{w}_u &= 0 \\ \text{For } v: \quad &-\boldsymbol{x}_v^\top Z + \frac{\gamma_{\mathcal{G}}}{\lambda_{\max}}\boldsymbol{L}_{\mathcal{G}}(\boldsymbol{w}_v - \boldsymbol{w}_u) + \lambda_1^{(g)}\operatorname{sign}(\boldsymbol{w}_v) + 2\lambda_2^{(g)}\boldsymbol{w}_v &= 0 \end{aligned} \tag{B.3}$$

We subtract the two equations, with $\operatorname{sign}(\boldsymbol{w}_u) = \operatorname{sign}(\boldsymbol{w}_v)$, which becomes:

$$(\boldsymbol{x}_v - \boldsymbol{x}_u)^\top Z + 2\left(\frac{\gamma_{\mathcal{G}}}{\lambda_{\max}} + \boldsymbol{L}_{\mathcal{G}}\lambda_2^{(g)}\right)(\boldsymbol{w}_u - \boldsymbol{w}_v) = 0 \tag{B.4}$$

Using the Cauchy-Schwarz inequality, we can see that :

$$(\boldsymbol{x}_v - \boldsymbol{x}_u)^\top Z \leq \|Z\|_2 \|\boldsymbol{x}_v - \boldsymbol{x}_u\|_2 \tag{B.5}$$

Using the equality $\|\boldsymbol{x}_v - \boldsymbol{x}_u\|_2 = \sqrt{\|x_v\|_2^2 + \|x_u\|_2^2 - 2x_v^\top x_u}$. And $x_v^\top x_u = \rho$ the estimated correlations between the variables u and v, we can show equation 6.7:

$$|\boldsymbol{w}_u^{(g)} - \boldsymbol{w}_v^{(g)}| \leq \frac{Q\sqrt{2(1-\rho)}}{2\left(\frac{\gamma_{\mathcal{G}}}{\lambda_{\max}} + \lambda_2^{(g)}\right)} \tag{B.6}$$

Note that $Q = \|Z\|_2$, a constant only dependent on the correlations of the other blocks. This concludes our proof.

Appendix C

Résumé en français

C.1 Introduction

Les tumeurs du système nerveux central (SNC) sont des néoplasmes situés dans le tissu cérébral ou dans la moelle épinière. Les gliomes sont d'un intérêt particulier. La cellule d'origine est une cellule gliale, c'est-à-dire une cellule non neurale qui ne produit pas d'impulsions électriques, du SNC. Les gliomes représentent environ 30% de toutes les tumeurs du SNC et 80% de toutes les tumeurs cérébrales malignes.

Le gliome infiltrant du tronc cérébral (Diffuse Intrinsic Pontine Glioma ou DIPG) est une tumeur cérébrale rare située dans le pons, principalement observée chez les enfants âgés de 5 à 7 ans. Elle est considérée comme l'une des tumeurs pédiatriques les plus agressives, avec un taux de survie inférieur à 10% au-delà des deux ans après le diagnostic (Fisher et al., 2000) et une médiane de survie globale inférieure à un an (Cohen et al., 2017). Le DIPG est classé comme un gliome diffus de la ligne médiane (DMG), principalement caractérisé par une mutation K27M des gènes codant pour la protéine histone H3 et/ou une perte de la tri-méthylation de H3K27. Des études récentes ouvrent des voies de stratifications des patients. Par exemple, suivant que la mutation K27M apparaît sur les sous-types d'histones H3.3 (H3F3A) ou H3.1 (HIST1H3B/C), la réponse clinique à la radiothérapie diffère (Castel et al., 2015). De même, il a été observée que de nombreux autres gènes peuvent se retrouver mutés dans le DIPG (entre autres PTEN ou P53) et que la surexpression d'un gène comme EZHIP peut entraîner la perte de triméthylation H3K27M (Castel et al., 2020). L'emplacement de la tumeur et ses altérations génomiques correspondantes fait du DIPG un type de tumeur complètement différent des autres tumeurs de haut grade comme celles de l'adulte.

Cependant, les connaissances d'aujourd'hui autour du DIPG ne sont pas suffisantes pour comprendre la tumeur, son évolution chez les patients ou le développement d'un traitement efficace. Actuellement, il n'existe aucun traitement curatif pour les patients atteints de DIPG, ce qui pousse à la recherche de nouvelles ressources et méthodes pour caractériser ces tumeurs par de nouveaux biomarqueurs.

Le dépistage par imagerie à résonance magnétique (IRM) a été intégré à la plupart des protocoles de diagnostic oncologique. Ces images sont devenues importantes parmi diverses techniques d'imagerie médicale en raison de leur sécurité et de leur richesse d'information. Ces dernières années ont connu de multiples avancées technologiques en imagerie biomédicale, apportant une meilleure qualité d'image

et une meilleure résolution spatiale. Ce qui a permis de caractériser les tumeurs de manière non invasive. L'analyse radiomique des IRM comprend toutes les étapes d'extraction et d'analyse d'un grand nombre de caractéristiques (plus de 200) d'une région d'intérêt (ROI) extraite des images cliniques. Ces caractéristiques décrivent la taille et la forme parmi d'autres propriétés visuelles apparentes des tumeurs. Puis des statistiques de premier ordre du signal d'intensité provenant des ROI sont calculées. Ensuite, les caractéristiques de statistiques de second ordre (texture) servent à décrire des configurations de niveau de gris fines et locales. L'hypothèse derrière l'analyse radiomique est que l'IRM peut capturer des caractéristiques de texture invisibles à l'œil nu, qui peuvent être associées à l'évolution d'une maladie.

Nous faisons l'hypothèse que combiner les caractéristiques radiomiques avec les données de biologie moléculaire permettra de construire des modèles riches et très informatifs sur la biologie des tumeurs cérébrales. L'objectif est de trouver de nouveaux biomarqueurs pronostiques hybrides, voire de caractériser la tumeur à partir des seuls descripteurs de l'imagerie, ce qui fournirait à terme un moyen totalement non-invasif de monitorer les DIPG ou les tumeurs cérébrales en général. Cette thèse est motivée par l'étude du DIPG inclue dans les études cliniques BIOMEDE dirigées par l'Institut Gustave Roussy. Ces gliomes pédiatriques sont rares (environ 40 nouveaux cas chaque année en France et 300 en Europe). Une fraction importante des données initialement prévue pour cette étude n'a pas été accessible. Nous avons décidé d'étendre les jeux de données étudiés aux gliomes de bas grade (LGG) et au glioblastome, pour lesquels des données multi-omiques sont publiquement disponibles grâce au projet The Cancer Genome Atlas (TCGA). Bien que ces tumeurs soient d'origine et de pronostic différents, nous utilisons ces données afin d'étudier les méthodes proposées et leurs limites.

Ce travail propose l'intégration des données d'imagerie avec les données génétiques afin de trouver des biomarqueurs. Dans un premier temps, nous nous intéressons à l'extraction des régions d'intérêt des images nécessaires pour une étude radiomique. Ensuite, nous proposons une procédure d'intégration des données multi-sources, qui prend en compte les graphes complexes d'interaction entre les gènes. Finalement, nous appliquons notre procédure sur les données disponibles afin de comparer ses performances avec d'autres modèles de la littérature et d'étudier l'apport de l'imagerie et du graphe aux données génétiques.

Le manuscrit de thèse est organisé en deux parties. La première partie est dédiée à la définition du contexte de l'étude. Elle débute par les définitions et notations utilisés dans toute la suite de la thèse (Chapitre 2). Ensuite, nous révisons les méthodes liées à notre étude (Chapitre 3). Enfin nous décrivons les données disponibles et utilisées durant la thèse avec les étapes de prétraitements utilisés (Chapitre 4). La deuxième partie de la thèse décrit notre contribution. Nous débutons par notre proposition pour la segmentation automatique de la tumeur (Chapitre 5). Ensuite, nous détaillons la méthode d'analyse multi-sources avec contrainte de graphe (Chapitre 6). Et enfin, nous appliquons nos méthodes sur les données disponibles (Chapitre 7). Dans la suite, nous résumons les trois chapitres de notre contribution.

C.2 Contribution

C.2.1 Chapitre 5 : Segmentation de la tumeur

La segmentation d'images est une étape clé pour l'analyse radiomique. L'identification de la région d'intérêt (ROI) de l'IRM est critique pour extraire des descripteurs significatifs de la tumeur en question. Étant donné un volume (Image) d'entrée d'une ou plusieurs modalités, la segmentation automatique des tumeurs cérébrales fait référence à la différenciation entre les voxels appartenant à la tumeur (et sa structure) et ceux appartenant au tissu cérébral sain. Pour cela, on utilise des modèles statistiques de classification appliqués à chaque voxel de l'image.

BraTS (Brain Tumour Segmentation challenge) est un défi international qui reflète l'avancement des méthodes de segmentation en oncologie. Depuis 2015, presque tous les modèles les plus performants du défi BraTS ont utilisé des réseaux de neurones convolutifs (Convolutionnal Neural Networks (CNN)). Notamment, l'architecture U-Net qui permet simultanément de reconstruire en sortie une segmentation de l'image d'origine. Nous proposons d'exploiter l'architecture U-Net (Ronneberger et al., 2015) en nous appuyant sur la détection d'objets.

Le défi consiste à obtenir des segmentations du DIPG sans entraîner des modèles sur cette tumeur car il n'y a pas assez de données disponibles, mais en les entraînant sur d'autres tumeurs pour lesquelles suffisamment de données sont publiquement accessibles. Nous avons décomposé le problème de délinéation des tumeurs en deux sous problèmes, i) de détection de tumeur dans des boîtes parallèlipidédiques (ou bounding-boxes) et ii) de segmentation des tumeurs dans ces boites. Nous proposons deux méthodes pour combiner un modèle de détection, YOU ONLY LOOK ONCE (YOLO) (Redmon et al., 2016), et le modèle UNET/BBUNET:

- La première méthode repose sur la suppression des segmentations en dehors de nos bounding-boxes, justifié par le fait que la probabilité qu'un voxel tumoral se trouve en dehors de la bounding-boxes est faible.
- La deuxième méthode positionne les bounding-boxes en entrée des modèles de segmentation. Le résultat de la segmentation est ensuite masqué comme dans la première proposition.

Nous avons étudié les segmentations obtenues par nos deux approches sur 3 jeux de données;

- TCGA-GBM, une cohorte de patients présentant glioblastome. Cette cohorte comporte la même tumeur, localisée dans la même région du cerveau que le jeu de données utilisé lors de l'entraînement.
- TCGA-LGG, une cohorte de patients avec un gliome de bas grade. Il s'agit de segmenter une tumeur différente, localisée dans la même région du cerveau que le jeu de données utilisé lors de l'entraînement.
- PREBIOMEDE, une petite cohorte privée de jeunes patients avec un DIPG et pour lesquels une segmentation experte est disponible.

Les résultats de nos propositions sont compétitifs par rapport à ceux des méthodes de l'état de l'art sur TCGA-GBM et meilleurs sur BRaTS LGG et PREBIOMEDE. De plus, notre méthode permet de proposer des segmentations robustes sur le DIPG ce qui évite d'avoir recours à un expert.

C.2.2 Chapitre 6 : Intégration multi-blocks

Dans le domaine de l'oncologie, des modèles statistiques sont utilisés pour la découverte de facteurs candidats qui influencent le développement de la pathologie ou sa réponse au traitement. Ceci est souvent réalisé en imposant des contraintes sur les modèles qui conduisent à une sélection des variables d'intérêt. Les contraintes basées sur des graphes a priori ont été utilisées dans la littérature comme un moyen d'améliorer la sélection des variables dans le modèle, ce qui fournit des modèles plus interprétables d'un point de vue biologique et/ou clinique. Cependant, les interactions entre les graphes choisis et le modèle sont mal caractérisées et on ignore comment ils impactent la sélection des variables. De plus, comme plusieurs graphes codant des informations différentes sont disponibles, on peut se demander comment le choix du graphe impacte les résultats obtenus.

Nous avons proposé d'étudier l'impact de la pénalité de graphe sur un modèle multi-blocs, dans le cadre de SGCCA (Sparse Generalised Canonical Correlation Analysis) (Tenenhaus et al., 2014). NetSGCCA consiste en une pénalité GraphNet ajoutée à un modèle SGCCA. Nous avons étudié l'effet de la pénalité sur le modèle à l'aide de l'ensemble de données TCGA-LGG.

Sur les données TCGA-LGG, en utilisant les données mRNA, miRNA et CNV, nous avons étudié l'effet de la normalisation du Laplacien du graphe dans la pénalité GraphNet. Ensuite, nous avons comparé différents graphes d'interaction gène-gène. Ces graphes sont différents dans leurs structures et informations contenues. Enfin, nous avons modifié un graphe en permutant ses nœuds et en supprimant quelques arêtes afin de vérifier l'utilité de l'information contenue dans le graphe.

Notre étude s'est focalisée, dans un premier temps, sur la sélection de variable avec la pénalité de graphe et sa stabilité. Dans un second temps, nous avons vérifié l'utilité des variables sélectionnées à prédire la survie et leur lien préalablement établi avec la tumeur étudiée dans la littérature.

Nous avons exhibé que cette pénalité permet de regrouper des variables, qui étaient séparées avec une pénalité l_1 classique, et nous avons observé une meilleure stabilité quand le Laplacien associé au graphe est normalisé. En comparant différents graphes, nous avons montré une corrélation positive entre la densité du graphe (définie comme le nombre d'arêtes) et le nombre de variables sélectionnées. Les résultats obtenus en utilisant les différents graphes ont été cohérents. Malgré un nombre de gènes sélectionnés différents avec les différents graphes, les ensembles de gènes sélectionnés étaient inclus les uns dans les autres. En permutant les nœuds du graphe, nous avons montré que la sémantique du graphe a un effet limité sur la sélection des variables, comparé à la densité du graphe.

Notre étude a aussi montré que les variables sélectionnées exhibent des voies biologiques précédemment associés aux gliomes de bas grade dans la littérature. Ceci démontre l'apport du graphe sur l'interprétabilité des résultats. De plus, la netSGCCA donne des résultats équivalents à la SGCCA en termes de prédiction de survie, voire dans certains cas meilleurs.

C.2.3 Chapitre 7 : Intégration Radio-génomique

Les caractéristiques radiomiques ont été utilisées avec succès pour différentes tâches et dans divers types de tumeurs. Cela comprend le DIPG, les glioblastomes et les gliomes de bas grade (LGG). Cependant, il n'y a pas de consensus sur la modalité d'image qui est la plus informative ou le type de caractéristique le plus important. De plus, la recherche a également montré que l'ajout de caractéristiques radiomiques aux données moléculaires peut améliorer la prédiction de la survie. Mais il

doit encore être intégré dans un cadre multibloc complet. Cela devrait nous aider à comprendre les interactions entre les caractéristiques radiomiques et génomiques.

Ce chapitre est consacré à l'application de netSGCCA pour la prédiction de survie lorsque des données génétiques sont associées à des caractéristiques radiomiques. Nous portons une attention particulière à l'ensemble des variables sélectionnées et à leur lien avec la maladie étudiée. Étant donné que le cadre multi-blocs avec lequel nous travaillons nécessite la disponibilité des données pour toutes les sources pour tous les patients, il en résulte naturellement moins d'échantillons disponibles pour une telle analyse. Par conséquent, nous établissons les performances de base de chaque bloc lorsque toutes les données disponibles sont prises en compte avant l'intégration. Ensuite, nous comparons l'approche mono-bloc avec l'approche multi-blocs. Nous comparons également la netSGCCA avec les modèles les plus performants parmi tous les travaux de l'état de l'art. En raison de l'indisponibilité des données de survie sur l'ensemble de données DIPG, cette étude a été limitée à l'ensemble de données TCGA-LGG.

La netSGCCA figurait parmi les modèles les plus performants de l'ensemble de validation et de test. Le C-index était stable pendant tous les essais, ce qui n'était pas le cas pour les autres modèles de Cox. Le modèle netSGCCA a également été capable de sélectionner des gènes et des voies déjà associés au développement des gliomes et au pronostic des LGG.

Les performances de base démontrent que les caractéristiques radiomiques peuvent être de bons prédicteurs de la survie du gliome de bas grade (LGG). La modalité T2w montre des performances particulièrement élevées, comme déjà observé dans la littérature. Malheureusement, nous n'avons pas pu reproduire ces résultats lors de l'utilisation de la radiomique sur les 83 patients ayant tous les blocs disponibles.

Dans l'ensemble, la restriction de la disponibilité des données a entravé les performances prédictives des blocs. Le modèle de Cox s'est montré également moins stable lorsque les données de seulement 83 patients ont été utilisées. De plus, la sélection de variables et les voies biologique identifiées dans l'étude de base et sur l'ensemble restreint ont donné des résultats significativement différents. Par conséquent, d'autres études sont nécessaires pour valider les résultats. Il est également nécessaire de réaliser cette étude sur un jeu de données plus important.

C.3 Conclusion

L'étude des tumeurs rares avec des modèles statistiques est difficile. Les données sont rares en raison de la nature de la tumeur. Cette rareté a deux conséquences. Des résultats peinent à être concluants et robustes et réclament des ensembles de données plus importants. La deuxième conséquence se traduit par des données collectées dans un contexte clinique, avec une grande variabilité de qualité et un nombre élevé de données manquantes. Tout en ayant conscience des difficultés que cela posent, un protocole plus strict conçu pour l'analyse statistique serait nécessaire.

Notre travail a proposé une approche de segmentation qui nous a permis d'obtenir des ROIs sur le DIPG. La poursuite de ces travaux pourrait porter sur une délimitation fine des tumeurs et de leurs sous-compartiments. De plus, la qualité de la segmentation et son impact sur les caractéristiques radiomiques doivent être évalués en détail.

Les données manquantes sont inhérentes aux études multi-blocs car toutes les données doivent être disponibles pour tous les patients et toutes les sources. A ce fait, s'ajoute dans le cas du DIPG, la collecte de données dans un contexte clinique qui est difficile à réaliser. Cette thèse n'a pas exploré la gestion des données manquantes dans l'intégration multi-blocs, ce qui a entraîné l'utilisation d'un ensemble de données relativement petit. Cela pourrait aussi constituer une direction nouvelle de recherche.

Enfin, ce travail propose l'utilisation de fonctionnalités radiomiques dans un cadre multi-blocs. Il faut cependant convenir, que les résultats obtenus ont pu être biaisées par les paramètres des séquences d'acquisition, les algorithmes de reconstruction et les méthodes de prétraitement. Des travaux supplémentaires considérant un vaste jeu de données permettant un contrôle strict de toutes les sources de variabilité permettrait de conclure plus avant sur les signatures d'image des tumeurs sans un contrôle strict de toutes les sources de variabilité.

Bibliography

Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, Sara Carvalho, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nature Communications*, 5(1):4006, Jun 2014. ISSN 2041-1723. doi: 10.1038/ncomms5006. URL https://doi.org/10.1038/ncomms5006. (Cited on page 3.)

Denisse Arcos-Montoya, Talia Wegman-Ostrosky, Sonia Mejía-Pérez, Marisol De la Fuente-Granada, Ignacio Camacho-Arroyo, Alejandro García-Carrancá, et al. Progesterone receptor together with PKCα expression as prognostic factors for astrocytomas malignancy. *Onco. Targets. Ther.*, 14: 3757–3768, June 2021. (Cited on page 110.)

Ricard Argelaguet, Britta Velten, Damien Arnol, Sascha Dietrich, Thorsten Zenz, John C Marioni, et al. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6), June 2018. doi: 10.15252/msb.20178124. URL https://doi.org/10.15252/msb.20178124. (Cited on page 30.)

Michael Ashburner, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, et al. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25(1):25–29, May 2000. doi: 10.1038/75556. URL https://doi.org/10.1038/75556. (Cited on page 34.)

Pierre Bady, Sylvain Dolédec, Bernard Dumont, and Jean-François Fruget. Multiple co-inertia analysis: a tool for assessing synchrony in the temporal variability of aquatic communities. *Comptes Rendus Biologies*, 327(1):29–36, 2004. ISSN 1631-0691. doi: https://doi.org/10.1016/j.crvi.2003.10.007. URL https://www.sciencedirect.com/science/article/pii/S1631069103003354. (Cited on page 30.)

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S. Kirby, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci. Data*, 4(1), September 2017. doi: 10.1038/sdata.2017.117. (Cited on page 39.)

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge, 2019. (Cited on page 39.)

G H Baltuch, N P Dooley, J G Villemure, and V W Yong. Protein kinase C and growth regulation of malignant gliomas. *Can. J. Neurol. Sci.*, 22(4):264–271, November 1995. (Cited on page 110.)

Philippe Bastien. Deviance residuals based pls regression for censored data in high dimensional setting. *Chemometrics and Intelligent Laboratory Systems*, 91(1):78–86, 2008. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2007.09.009. URL https://www.sciencedirect.com/science/article/pii/S0169743907001931. Selected papers presented at the Chemometrics Congress "CHIMIOMETRIE 2006" Paris, France, 30 November - 1 December 2006. (Cited on pages 81 and 93.)

H.H. Bauschke and J.M. Borwein. Dykstra's alternating projection algorithm for two sets. *Journal of Approximation Theory*, 79(3):418–443, 1994. ISSN 0021-9045. doi: https://doi.org/10.1006/jath.1994.1136. URL https://www.sciencedirect.com/science/article/pii/S0021904584711361. (Cited on pages 75 and 76.)

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009. (Cited on page 74.)

Mostefa Ben naceur, Mohamed Akil, Rachida Saouli, and Rostom Kachouri. Fully automatic brain tumor segmentation with deep learning-based selective attention using overlapping patches and multi-class weighted cross-entropy. *Medical Image Analysis*, 63:101692, 2020. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2020.101692. URL https://www.sciencedirect.com/science/article/pii/S1361841520300578. (Cited on page 27.)

Howard D. Bondell and Brian J. Reich. Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with OSCAR. *Biometrics*, 64(1):115–123, June 2007. doi: 10.1111/j.1541-0420.2007.00843.x. URL https://doi.org/10.1111/j.1541-0420.2007.00843.x. (Cited on page 33.)

Anne-Laure Boulesteix, Riccardo De Bin, Xiaoyu Jiang, and Mathias Fuchs. IPF-LASSO: Integrative L 1 -Penalized Regression with Penalty Factors for Prediction Based on Multi-Omics Data. *Computational and Mathematical Methods in Medicine*, 2017:1–14, 2017. ISSN 1748-670X, 1748-6718. doi: 10.1155/2017/7691937. URL https://www.hindawi.com/journals/cmmm/2017/7691937/. (Cited on page 31.)

Cameron W. Brennan, Roel G.W. Verhaak, Aaron McKenna, Benito Campos, Houtan Noushmehr, Sofie R. Salama, et al. The somatic genomic landscape of glioblastoma. *Cell*, 155(2):462–477, October 2013. doi: 10.1016/j.cell.2013.09.034. URL https://doi.org/10.1016/j.cell.2013.09.034. (Cited on pages 1 and 39.)

R.P. Brent. *Algorithms for minimization without derivatives*. Prentice-Hall, 1973. ISBN 0-13-022335-2. (Cited on page 54.)

N. E. Breslow. Analysis of survival data under the proportional hazards model. *International Statistical Review / Revue Internationale de Statistique*, 43(1):45–57, 1975. ISSN 03067734, 17515823. URL http://www.jstor.org/stable/1402659. (Cited on page 18.)

Patrik Brynolfsson, David Nilsson, Turid Torheim, Thomas Asklund, Camilla Thellenberg Karlsson, Johan Trygg, et al. Haralick texture features from apparent diffusion coefficient (adc) mri images depend on imaging and pre-processing parameters. *Scientific Reports*, 7(1):4041, Jun 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-04151-4. URL https://doi.org/10.1038/s41598-017-04151-4. (Cited on page 25.)

Stanislaw R. Burzynski, Tomasz J. Janicki, Gregory S. Burzynski, and Ania Marszalek. The response and survival of children with recurrent diffuse intrinsic pontine glioma based on phase ii study of antineoplastons a10 and as2-1 in patients with brainstem glioma. *Child's Nervous System*, 30(12): 2051–2061, Dec 2014. ISSN 1433-0350. doi: 10.1007/s00381-014-2401-z. URL https://doi.org/10.1007/s00381-014-2401-z. (Cited on page 2.)

Raphael Calmon, Stephanie Puget, Pascale Varlet, Kevin Beccaria, Thomas Blauwblomme, David Grevent, et al. Multimodal magnetic resonance imaging of treatment-induced changes to diffuse infiltrating pontine gliomas in children and correlation to patient progression-free survival. *International Journal of Radiation Oncology Biology Physics*, 99(2):476–485, October 2017. doi: 10.1016/j.ijrobp.2017.04.007. (Cited on page 41.)

Raphaël Calmon, Volodia Dangouloff-Ros, Pascale Varlet, Christophe Deroulers, Cathy Philippe, Marie-Anne Debily, et al. Radiogenomics of diffuse intrinsic pontine gliomas (DIPGs): correlation of histological and biological characteristics with multimodal MRI features. *European Radiology*, May 2021. doi: 10.1007/s00330-021-07991-x. (Cited on page 41.)

Angus J. Cameron, Katarzyna J. Procyk, Michael Leitges, and Peter J. Parker. Pkc alpha protein but not kinase activity is critical for glioma cell proliferation and survival. *International Journal of Cancer*, 123(4):769–779, 2008. doi: https://doi.org/10.1002/ijc.23560. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ijc.23560. (Cited on page 110.)

Laura Cantini, Pooya Zakeri, Celine Hernandez, Aurelien Naldi, Denis Thieffry, Elisabeth Remy, and Anaïs Baudot. Benchmarking joint multi-omics dimensionality reduction approaches for the study of cancer. *Nature Communications*, 12(1), January 2021. doi: 10.1038/s41467-020-20430-7. URL https://doi.org/10.1038/s41467-020-20430-7. (Cited on pages 3 and 31.)

Adrià Casamitjana, Santi Puch, Asier Aduriz, and Verónica Vilaplana. 3d convolutional neural networks for brain tumor segmentation: A comparison of multi-resolution architectures. In *BrainLes@MICCAI*, 2016. (Cited on page 28.)

David Castel, Cathy Philippe, Raphaël Calmon, Ludivine Le Dret, Nathalène Truffaux, Nathalie Boddaert, et al. Histone h3f3a and HIST1h3b k27m mutations define two subgroups of diffuse intrinsic pontine gliomas with different prognosis and phenotypes. *Acta Neuropathol.*, 130(6):815–827, September 2015. doi: 10.1007/s00401-015-1478-0. (Cited on pages 2, 65, and 121.)

David Castel, Thomas Kergrohen, Arnault Tauziède-Espariat, Alan Mackay, Samia Ghermaoui, Emmanuèle Lechapt, et al. Histone h3 wild-type dipg/dmg overexpressing ezhip extend the spectrum diffuse midline gliomas with prc2 inhibition beyond h3-k27m mutation. *Acta Neuropatho-*

logica, 139(6):1109–1113, Jun 2020. ISSN 1432-0533. doi: 10.1007/s00401-020-02142-w. URL https://doi.org/10.1007/s00401-020-02142-w. (Cited on pages 37 and 121.)

Renee Cattell, Shenglan Chen, and Chuan Huang. Robustness of radiomic features in magnetic resonance imaging: review and a phantom study. *Visual Computing for Industry, Biomedicine, and Art*, 2(1):19, Nov 2019. ISSN 2524-4442. doi: 10.1186/s42492-019-0025-6. URL https://doi.org/10.1186/s42492-019-0025-6. (Cited on page 26.)

E. G. Cerami, B. E. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, et al. Pathway commons, a web resource for biological pathway data. *Nucleic Acids Research*, 39(Database):D685–D690, November 2010. doi: 10.1093/nar/gkq1039. URL https://doi.org/10.1093/nar/gkq1039. (Cited on page 81.)

Hamza Chegraoui, Cathy Philippe, Volodia Dangouloff-Ros, Antoine Grigis, Raphael Calmon, Nathalie Boddaert, et al. Object detection improves tumour segmentation in mr images of rare brain tumours. *Cancers*, 13(23), 2021a. ISSN 2072-6694. doi: 10.3390/cancers13236113. URL https://www.mdpi.com/2072-6694/13/23/6113. (Cited on pages 6 and 49.)

Hamza Chegraoui, Amine Rebei, Cathy Philippe, and Vincent Frouin. Prediction performance of radiomic features when obtained using an object detection framework. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1382–1385, 2021b. doi: 10.1109/ISBI48211.2021.9434148. (Cited on page 67.)

Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, et al. Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics*, 14(1), April 2013. doi: 10.1186/1471-2105-14-128. URL https://doi.org/10.1186/1471-2105-14-128. (Cited on pages 88 and 94.)

Kenneth J. Cohen, Nada Jabado, and Jacques Grill. Diffuse intrinsic pontine gliomas—current management and new biologic insights. is there a glimmer of hope? *Neuro-Oncol.*, 19(8):1025–1034, March 2017. doi: 10.1093/neuonc/nox021. (Cited on pages 2 and 121.)

The Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, 12 2020. ISSN 0305-1048. doi: 10.1093/nar/gkaa1113. URL https://doi.org/10.1093/nar/gkaa1113. (Cited on page 34.)

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995. (Cited on page 16.)

D. R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202, 1972. doi: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1972.tb00899.x. (Cited on page 17.)

Antonio Criminisi, Darko Zikic, Ben Glocker, and Jamie Shotton. Context-sensitive classification forests for segmentation of brain tumor tissues. In *MICCAI 2012 Challenge on Multimodal Brain Tumor Segmentation*, October 2012. URL https://www.microsoft.com/en-us/research/publication/

context-sensitive-classification-forests-for-segmentation-of-brain-tumor-tissues/. (Cited on page 26.)

Amicie de Pierrefeu, Thomas Fovet, Fouad Hadj-Selem, Tommy Löfstedt, Philippe Ciuciu, Stephanie Lefebvre, et al. Prediction of activation patterns preceding hallucinations in patients with schizophrenia using machine learning with structured sparsity. *Human brain mapping*, 39(4):1777–1788, April 2018. doi: 10.1002/hbm.23953. URL http://doi.wiley.com/10.1002/hbm.23953. (Cited on page 76.)

Lei Du, Heng Huang, Jingwen Yan, Sungeun Kim, Shannon L. Risacher, Mark Inlow, et al. Structured sparse canonical correlation analysis for brain imaging genetics: an improved GraphNet method. *Bioinformatics*, 32(10):1544–1551, May 2016. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btw033. URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw033. (Cited on page 34.)

Lei Du, Kefei Liu, Xiaohui Yao, Shannon L. Risacher, Junwei Han, Andrew J. Saykin, et al. Detecting genetic associations with brain imaging phenotypes in Alzheimer's disease via a novel structured SCCA approach. *Medical Image Analysis*, 61:101656, April 2020. ISSN 1361-8415. doi: 10.1016/j.media.2020.101656. URL https://www.sciencedirect.com/science/article/pii/S1361841520300232. (Cited on pages 34 and 76.)

Pavel Dvorak and B. Menze. Structured prediction with convolutional neural networks for multimodal brain tumor segmentation. pages 13–24, 10 2015. (Cited on page 26.)

B. Escofier and J. Pagès. Multiple factor analysis (afmult package). *Computational Statistics & Data Analysis*, 18(1):121–140, 1994. ISSN 0167-9473. doi: https://doi.org/10.1016/0167-9473(94)90135-X. URL https://www.sciencedirect.com/science/article/pii/016794739490135X. (Cited on page 30.)

Paul G. Fisher, Steven N. Breiter, Benjamin S. Carson, Moody D. Wharam, Jeffery A. Williams, Jon D. Weingart, et al. A clinicopathologic reappraisal of brain stem tumor classification. *Cancer*, 89(7):1569–1576, October 2000. doi: 10.1002/1097-0142(20001001)89:7<1569::aid-cncr22>3.0.co;2-0. (Cited on pages 2 and 121.)

Vladimir Fonov, Alan C. Evans, Kelly Botteron, C. Robert Almli, Robert C. McKinstry, and D. Louis Collins. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1):313–327, 2011. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2010.07.033. (Cited on page 57.)

Deborah A. Forst, Brian V. Nahed, Jay S. Loeffler, and Tracy T. Batchelor. Low-Grade Gliomas. *The Oncologist*, 19(4):403–413, 03 2014. ISSN 1083-7159. doi: 10.1634/theoncologist.2013-0345. URL https://doi.org/10.1634/theoncologist.2013-0345. (Cited on page 1.)

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. doi: 10.18637/jss.v033.i01. URL https://www.jstatsoft.org/v33/i01/. (Cited on page 96.)

F Gaillard and J Yap. Diffuse glioma. 2017. URL https://radiopaedia.org/articles/52485. (Cited on page 2.)

Matthew Gallitto, Stanislav Lazarev, Isaac Wasserman, James M Stafford, Suzanne L Wolden, Stephanie A Terezakis, et al. Role of radiation therapy in the management of diffuse intrinsic pontine glioma: A systematic review. *Adv. Radiat. Oncol.*, 4(3):520–531, July 2019. (Cited on page 2.)

Floyd H. Gilles, C. Jane Tavaré, E. Becker Laurence, Peter C. Burger, Allan J. Yates, Ian F. Pollack, and Jonathan L. Finlay. Pathologist interobserver variability of histologic features in childhood brain tumors: Results from the ccg-945 study. *Pediatric and Developmental Pathology*, 11(2):108–117, 2008. doi: 10.2350/07-06-0303.1. URL https://doi.org/10.2350/07-06-0303.1. (Cited on page 1.)

McKinsey L. Goodenberger and Robert B. Jenkins. Genetics of adult glioma. *Cancer Genetics*, 205 (12):613–621, December 2012. doi: 10.1016/j.cancergen.2012.10.009. URL https://doi.org/10.1016/j.cancergen.2012.10.009. (Cited on page 1.)

Jessica Goya-Outi, Fanny Orlhac, Raphael Calmon, Agusti Alentorn, Christophe Nioche, Cathy Philippe, et al. Computation of reliable textural indices from multimodal brain mri: suggestions based on a study of patients with diffuse intrinsic pontine glioma. *Physics in Medicine & Biology*, 63(10):105003, may 2018. doi: 10.1088/1361-6560/aabd21. URL https://dx.doi.org/10.1088/1361-6560/aabd21. (Cited on page 25.)

Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E. Taylor. Interpretable whole-brain prediction analysis with graphnet. *NeuroImage*, 72:304–321, 2013. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2012.12.062. URL https://www.sciencedirect.com/science/article/pii/S1053811912012487. (Cited on page 33.)

Qian Guan, Li Yuan, Ao Lin, Huiran Lin, Xiaokai Huang, Jichen Ruan, and Zhenjian Zhuo. KRAS gene polymorphisms are associated with the risk of glioma: a two-center case-control study. *Transl. Pediatr.*, 10(3):579–586, March 2021. (Cited on page 110.)

N. Guigui, C. Philippe, A. Gloaguen, S. Karkar, V. Guillemot, T. Löfstedt, and V. Frouin. Network regularization in imaging genetics improves prediction performances and model interpretability on alzheimer's disease. In *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pages 1403–1406, 2019. doi: 10.1109/ISBI.2019.8759593. (Cited on page 72.)

Deliang Guo, Erica Hlavin Bell, and Arnab Chakravarti. Lipid metabolism emerges as a promising target for malignant glioma therapy. *CNS Oncology*, 2(3):289–299, May 2013. doi: 10.2217/cns.13.20. URL https://doi.org/10.2217/cns.13.20. (Cited on page 106.)

Nalin Gupta, Liliana C Goumnerova, Peter Manley, Susan N Chi, Donna Neuberg, Maneka Puligandla, et al. Prospective feasibility and safety assessment of surgical biopsy for patients with newly diagnosed diffuse intrinsic pontine glioma. *Neuro-Oncology*, 20(11):1547–1555, 05 2018. ISSN 1522-8517. doi: 10.1093/neuonc/noy070. URL https://doi.org/10.1093/neuonc/noy070. (Cited on page 2.)

Ho-Shin Gwak and Hyeon Jin Park. Developing chemotherapy for diffuse pontine intrinsic gliomas (dipg). *Critical Reviews in Oncology/Hematology*, 120:111–119, 2017. ISSN 1040-8428. doi: https://doi.org/10.1016/j.critrevonc.2017.10.013. URL https://www.sciencedirect.com/science/article/pii/S1040842817300653. (Cited on page 2.)

Jianfeng Han, Christopher A Alvarez-Breckenridge, Qi-En Wang, and Jianhua Yu. TGF-β signaling and its targeting for glioma treatment. *Am. J. Cancer Res.*, 5(3):945–955, February 2015. (Cited on page 88.)

Todd C. Hankinson, Elizabeth J. Campagna, Nich Olas K. Foreman, and Michael H. Handler. Interpretation of magnetic resonance images in diffuse intrinsic pontine glioma: A survey of pediatric neurosurgeons - Clinical article. *Journal of Neurosurgery: Pediatrics*, 8(1):97–102, 2011. ISSN 19330707. doi: 10.3171/2011.4.PEDS1180. (Cited on page 51.)

Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, et al. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.05.004. URL https://www.sciencedirect.com/science/article/pii/S1361841516300330. (Cited on page 27.)

Moritz Herrmann, Philipp Probst, Roman Hornung, Vindi Jurinovic, and Anne-Laure Boulesteix. Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3), August 2020. doi: 10.1093/bib/bbaa167. URL https://doi.org/10.1093/bib/bbaa167. (Cited on pages 3, 32, 40, 89, and 93.)

Nicholas J. Higham. Computing the nearest correlation matrix—a problem from finance. *IMA Journal of Numerical Analysis*, 22(3):329–343, 2002. doi: 10.1093/imanum/22.3.329. (Cited on page 78.)

Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE, 1995. (Cited on page 18.)

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. doi: 10.1080/00401706.1970.10488634. URL https://www.tandfonline.com/doi/abs/10.1080/00401706.1970.10488634. (Cited on page 16.)

Matan Hofree, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115, Nov 2013. ISSN 1548-7105. doi: 10.1038/nmeth.2651. URL https://doi.org/10.1038/nmeth.2651. (Cited on pages 5 and 46.)

Roman Hornung and Marvin N. Wright. Block forests: random forests for blocks of clinical and omics covariate data. *BMC Bioinformatics*, 20(1):358, Jun 2019. ISSN 1471-2105. doi: 10.1186/s12859-019-2942-y. URL https://doi.org/10.1186/s12859-019-2942-y. (Cited on pages 32 and 93.)

Paul Horst. Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17(4):331–347, 1961. doi: https://doi.org/10.1002/1097-4679(196110)17:4<331::AID-JCLP2270170402>3.0.CO;2-D. (Cited on page 30.)

HAROLD HOTELLING. RELATIONS BETWEEN TWO SETS OF VARIATES*. *Biometrika*, 28 (3-4):321–377, 12 1936. ISSN 0006-3444. doi: 10.1093/biomet/28.3-4.321. URL https://doi.org/10.1093/biomet/28.3-4.321. (Cited on page 19.)

Rui Hua, Quan Huo, Yaozong Gao, He Sui, Bing Zhang, Yu Sun, et al. Segmenting brain tumor using cascaded v-nets in multimodal mr images. *Frontiers in Computational Neuroscience*, 14, 2020. ISSN 1662-5188. doi: 10.3389/fncom.2020.00009. URL https://www.frontiersin.org/articles/10.3389/fncom.2020.00009. (Cited on page 28.)

Jiaqi Huang, Stephanie J. Weinstein, Cari M. Kitahara, Edward D. Karoly, Joshua N. Sampson, and Demetrius Albanes. A prospective study of serum metabolites and glioma risk. *Oncotarget*, 8 (41):70366–70377, July 2017. doi: 10.18632/oncotarget.19705. URL https://doi.org/10.18632/oncotarget.19705. (Cited on page 106.)

David M. Irvin, Robert S. McNeill, Ryan E. Bash, and C. Ryan Miller. Intrinsic astrocyte heterogeneity influences tumor growth in glioma mouse models. *Brain Pathology*, 27(1):36–50, 2017. doi: https://doi.org/10.1111/bpa.12348. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/bpa.12348. (Cited on page 88.)

Fabian Isensee, Philipp Kickingereder, Wolfgang Wick, Martin Bendszus, and Klaus H. Maier-Hein. Brain tumor segmentation and radiomics survival prediction: Contribution to the brats 2017 challenge. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Bjoern Menze, and Mauricio Reyes, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 287–297, Cham, 2018. Springer International Publishing. ISBN 978-3-319-75238-9. (Cited on page 28.)

Fabian Isensee, Paul F. Jaeger, Simon A. A. 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, Feb 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01008-z. (Cited on page 67.)

Atiq Islam, Syed M. S. Reza, and Khan M. Iftekharuddin. Multifractal texture estimation for detection and segmentation of brain tumors. *IEEE Transactions on Biomedical Engineering*, 60(11):3204–3215, 2013. doi: 10.1109/TBME.2013.2271383. (Cited on page 26.)

Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical Image Analysis*, 5(2):143–156, 2001. ISSN 1361-8415. doi: https://doi.org/10.1016/S1361-8415(01)00036-6. URL https://www.sciencedirect.com/science/article/pii/S1361841501000366. (Cited on page 44.)

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *NeuroImage*, 17 (2):825–841, 2002. ISSN 1053-8119. doi: https://doi.org/10.1006/nimg.2002.1132. URL https://www.sciencedirect.com/science/article/pii/S1053811902911328. (Cited on page 44.)

Glenn Jocher, Yonghye Kwon, Guigarfr, Josh Veitch-Michaelis, Perry0418, Ttayu, et al. ultralytics/yolov3: 43.1map@0.5:0.95 on coco2014, 2020. (Cited on page 52.)

Glenn Jocher, Alex Stoken, Jirka Borovec, NanoCode012, ChristopherSTAN, Liu Changyu, et al. ultralytics/yolov5: v4.0 - nn.silu() activations, weights & biases logging, pytorch hub integration, 2021. (Cited on page 55.)

Kjetil Jørgensen, Bjørn-Helge Mevik, and Tormod Næs. Combining designed experiments with several blocks of spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*, 88(2):154–166, 2007. ISSN 0169-7439. doi: https://doi.org/10.1016/j.chemolab.2007.04.002. URL https://www.sciencedirect.com/science/article/pii/S0169743907000688. (Cited on page 31.)

Konstantinos Kamnitsas, Enzo Ferrante, Sarah Parisot, Christian Ledig, Aditya V. Nori, Antonio Criminisi, et al. Deepmedic for brain tumor segmentation. In Alessandro Crimi, Bjoern Menze, Oskar Maier, Mauricio Reyes, Stefan Winzeck, and Heinz Handels, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 138–149, Cham, 2016. Springer International Publishing. ISBN 978-3-319-55524-9. (Cited on page 27.)

Konstantinos Kamnitsas, Christian Ledig, Virginia F.J. Newcombe, Joanna P. Simpson, Andrew D. Kane, David K. Menon, et al. Efficient multi-scale 3d CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61–78, February 2017. doi: 10.1016/j.media.2016.10.004. (Cited on pages x, 56, and 60.)

M. Kanehisa. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30, January 2000. doi: 10.1093/nar/28.1.27. URL https://doi.org/10.1093/nar/28.1.27. (Cited on pages 34 and 81.)

Amanda M. Katz, Nduka M. Amankulor, Ken Pitter, Karim Helmy, Massimo Squatrito, and Eric C. Holland. Astrocyte-specific expression patterns associated with the PDGF-induced glioma microenvironment. *PLoS ONE*, 7(2):e32453, February 2012. doi: 10.1371/journal.pone.0032453. URL https://doi.org/10.1371/journal.pone.0032453. (Cited on page 88.)

Rejin Kebudi, Fatma Betul Cakir, Sema Buyukkapu Bay, Omer Gorgun, Pelin Altınok, Ayça Iribas, et al. Nimotuzumab-containing regimen for pediatric diffuse intrinsic pontine gliomas: a retrospective multicenter study and review of the literature. *Child's Nervous System*, 35(1):83–89, Jan 2019. ISSN 1433-0350. doi: 10.1007/s00381-018-4001-9. URL https://doi.org/10.1007/s00381-018-4001-9. (Cited on page 2.)

J. R. KETTENRING. Canonical analysis of several sets of variables. *Biometrika*, 58(3):433–451, 12 1971. ISSN 0006-3444. doi: 10.1093/biomet/58.3.433. URL https://doi.org/10.1093/biomet/58.3.433. (Cited on page 30.)

Philipp Kickingereder, Sina Burth, Antje Wick, Michael Götz, Oliver Eidel, Heinz-Peter Schlemmer, et al. Radiomic profiling of glioblastoma: Identifying an imaging predictor of patient survival with improved performance over established clinical and radiologic risk models. *Radiology*, 280(3):880–889, 2016. doi: 10.1148/radiol.2016160845. URL https://doi.org/10.1148/radiol.2016160845. PMID: 27326665. (Cited on pages 23 and 65.)

Philipp Kickingereder, Ulf Neuberger, David Bonekamp, Paula L Piechotta, Michael Götz, Antje Wick, et al. Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma. *Neuro-Oncology*, 20(6):848–857, 09 2017. ISSN 1522-8517. doi: 10.1093/neuonc/nox188. URL https://doi.org/10.1093/neuonc/nox188. (Cited on page 23.)

Jung-Ae Kim. Peroxisome metabolism in cancer. *Cells*, 9(7):1692, July 2020. doi: 10.3390/cells9071692. URL https://doi.org/10.3390/cells9071692. (Cited on page 106.)

Mansu Kim, Ji Sun Kim, Jinyoung Youn, Hyunjin Park, and Jin Whan Cho. GraphNet-based imaging biomarker model to explain levodopa-induced dyskinesia in Parkinson's disease. *Computer Methods and Programs in Biomedicine*, 196:105713, November 2020. ISSN 0169-2607. doi: 10.1016/j.cmpb.2020.105713. URL https://www.sciencedirect.com/science/article/pii/S0169260720315467. (Cited on page 34.)

Yoo-Ah Kim, Stefan Wuchty, and Teresa M. Przytycka. Identifying causal genes and dysregulated pathways in complex diseases. *PLOS Computational Biology*, 7(3):1–13, 03 2011. doi: 10.1371/journal.pcbi.1001095. URL https://doi.org/10.1371/journal.pcbi.1001095. (Cited on page 5.)

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. (Cited on page 56.)

Simon Klau, Vindi Jurinovic, Roman Hornung, Tobias Herold, and Anne-Laure Boulesteix. Priority-lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*, 19(1), September 2018. doi: 10.1186/s12859-018-2344-6. URL https://doi.org/10.1186/s12859-018-2344-6. (Cited on pages 31 and 93.)

Simon Klau, Roman Hornung, and Alina Bauer. *prioritylasso: Analyzing Multiple Omics Data with an Offset Approach*, 2020. URL https://CRAN.R-project.org/package=prioritylasso. R package version 0.2.5. (Cited on page 93.)

Christiane B. Knobbe, Julia Reifenberger, and Guido Reifenberger. Mutation analysis of the ras pathway genes NRAS, HRAS, KRAS and BRAF in glioblastomas. *Acta Neuropathologica*, 108 (6):467–470, October 2004. doi: 10.1007/s00401-004-0929-9. URL https://doi.org/10.1007/s00401-004-0929-9. (Cited on page 88.)

Ziren Kong, Yusong Lin, Chendan Jiang, Longfei Li, Zehua Liu, Yuekun Wang, et al. 18f-fdg-pet-based radiomics signature predicts mgmt promoter methylation status in primary diffuse glioma. *Cancer Imaging*, 19(1):58, Aug 2019. ISSN 1470-7330. doi: 10.1186/s40644-019-0246-0. URL https://doi.org/10.1186/s40644-019-0246-0. (Cited on page 23.)

Richard Kramer. *Chemometric Techniques for Quantitative Analysis*. CRC Press, June 1998. doi: 10.1201/9780203909805. URL https://doi.org/10.1201/9780203909805. (Cited on page 16.)

Giedre Krenciute, Brooke L. Prinzing, Zhongzhen Yi, Meng-Fen Wu, Hao Liu, Gianpietro Dotti, et al. Transgenic Expression of IL15 Improves Antiglioma Activity of IL13Rα2-CAR T Cells but Results

in Antigen Loss Variants. *Cancer Immunology Research*, 5(7):571–581, 07 2017. ISSN 2326-6066. doi: 10.1158/2326-6066.CIR-16-0376. URL https://doi.org/10.1158/2326-6066.CIR-16-0376. (Cited on page 110.)

Maxim V. Kuleshov, Matthew R. Jones, Andrew D. Rouillard, Nicolas F. Fernandez, Qiaonan Duan, Zichen Wang, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Research*, 44(W1):W90–W97, May 2016. doi: 10.1093/nar/gkw377. URL https://doi.org/10.1093/nar/gkw377. (Cited on pages 88 and 94.)

Michael S. Lawrence, Petar Stojanov, Craig H. Mermel, James T. Robinson, Levi A. Garraway, Todd R. Golub, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, 505(7484):495–501, January 2014. doi: 10.1038/nature12912. URL https://doi.org/10.1038/nature12912. (Cited on page 5.)

Marine Le Morvan, Andrei Zinovyev, and Jean-Philippe Vert. Netnorm: Capturing cancer-relevant information in somatic exome mutation data with gene networks for cancer stratification and prognosis. *PLOS Computational Biology*, 13(6):1–29, 06 2017. doi: 10.1371/journal.pcbi.1005573. URL https://doi.org/10.1371/journal.pcbi.1005573. (Cited on page 5.)

Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, Oct 1999. ISSN 1476-4687. doi: 10.1038/44565. URL https://doi.org/10.1038/44565. (Cited on page 18.)

Dongjin Leng, Linyi Zheng, Yuqi Wen, Yunhao Zhang, Lianlian Wu, Jing Wang, et al. A benchmark study of deep learning-based multi-omics data fusion methods for cancer. *Genome Biology*, 23(1):171, Aug 2022. ISSN 1474-760X. doi: 10.1186/s13059-022-02739-2. URL https://doi.org/10.1186/s13059-022-02739-2. (Cited on page 32.)

Thomas Leung and Jitendra Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1):29–44, Jun 2001. ISSN 1573-1405. doi: 10.1023/A:1011126920638. URL https://doi.org/10.1023/A:1011126920638. (Cited on page 26.)

Caiyan Li and Hongzhe Li. Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics*, 24(9):1175–1182, 03 2008. ISSN 1367-4803. doi: 10.1093/bioinformatics/btn081. URL https://doi.org/10.1093/bioinformatics/btn081. (Cited on pages 5, 33, 74, 76, and 119.)

A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdottir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, 27(12):1739–1740, May 2011. doi: 10.1093/bioinformatics/btr260. URL https://doi.org/10.1093/bioinformatics/btr260. (Cited on page 81.)

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, et al. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1. (Cited on page 55.)

Katherine T Lind, Hannah V Chatwin, John DeSisto, Philip Coleman, Bridget Sanford, Andrew M Donson, et al. Novel RAF fusions in pediatric low-grade gliomas demonstrate MAPK pathway activation. *Journal of Neuropathologýs Experimental Neurology*, 80(12):1099–1107, November 2021. doi: 10.1093/jnen/nlab110. URL https://doi.org/10.1093/jnen/nlab110. (Cited on page 88.)

Mingfa Liu, Zhennan Xu, Zepeng Du, Bingli Wu, Tao Jin, Ke Xu, et al. The identification of key genes and pathways in glioma by bioinformatics analysis. *Journal of Immunology Research*, 2017: 1278081, Dec 2017. ISSN 2314-8861. doi: 10.1155/2017/1278081. URL https://doi.org/10.1155/2017/1278081. (Cited on page 101.)

Xing Liu, Yiming Li, Zenghui Qian, Zhiyan Sun, Kaibin Xu, Kai Wang, et al. A radiomic signature as a non-invasive predictor of progression-free survival in patients with lower-grade gliomas. *NeuroImage: Clinical*, 20:1070–1077, 2018. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2018.10.014. URL https://www.sciencedirect.com/science/article/pii/S2213158218303243. (Cited on pages 23 and 112.)

Eric F Lock, Katherine A Hoadley, J S Marron, and Andrew B Nobel. Joint and individual variation explained (jive) for integrated analysis of multiple data types. *Ann. Appl. Stat.*, 7(1):523–542, March 2013. (Cited on page 30.)

Tommy Löfstedt, Fouad Hadj-Selem, Vincent Guillemot, Cathy Philippe, Edouard Duchesnay, Vincent Frouin, and Arthur Tenenhaus. Structured variable selection for regularized generalized canonical correlation analysis. In *Springer Proceedings in Mathematicsś Statistics*, pages 129–139. Springer International Publishing, 2016. doi: 10.1007/978-3-319-40643-5_10. URL https://doi.org/10.1007/978-3-319-40643-5_10. (Cited on page 74.)

Tommy Lofstedt, Fouad Hadj-Selem, Vincent Guillemot, Cathy Philippe, Edouard Duchesnay, Vincent Frouin, and Arthur Tenenhaus. Structured {Variable} {Selection} for {Regularized} {Generalized} {Canonical} {Correlation} {Analysis}, {The} {Multiple} {Facets} of {Partial} {Least} {Squares} and {Related} {Methods}. In *Springer {Proceedings} in {Mathematics} {\textbackslash}& {Statistics}*, The {Multiple} {Facets} of {Partial} {Least} {Squares} and {Related} {Methods}, pages 129–139. 2016. URL https://hal-centralesupelec.archives-ouvertes.fr/hal-01396614. (Cited on page 76.)

Fang Longwei and He Huiguang. Three pathways u-net for brain tumor segmentation. In *BrainLes@MICCAI*, 2018. (Cited on page 29.)

David N. Louis, Arie Perry, Guido Reifenberger, Andreas von Deimling, Dominique Figarella-Branger, Webster K. Cavenee, et al. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta Neuropathologica*, 131(6):803–820, May 2016. doi: 10.1007/s00401-016-1545-1. URL https://doi.org/10.1007/s00401-016-1545-1. (Cited on pages 1 and 2.)

David N Louis, Arie Perry, Pieter Wesseling, Daniel J Brat, Ian A Cree, Dominique Figarella-Branger, et al. The 2021 WHO classification of tumors of the central nervous system: a summary. *Neuro-*

Oncology, 23(8):1231–1251, June 2021. doi: 10.1093/neuonc/noab106. URL https://doi.org/10.1093/neuonc/noab106. (Cited on pages 1 and 2.)

Nauman Malik, Benjamin Geraghty, Archya Dasgupta, Pejman Jabehdar Maralani, Michael Sandhu, Jay Detsky, et al. Mri radiomics to differentiate between low grade glioma and glioblastoma peritumoral region. *Journal of Neuro-Oncology*, 155(2):181–191, Nov 2021. ISSN 1573-7373. doi: 10.1007/s11060-021-03866-9. URL https://doi.org/10.1007/s11060-021-03866-9. (Cited on page 23.)

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943. (Cited on page 16.)

Roger McLendon, Allan Friedman, Darrell Bigner, Erwin G. Van Meir, Daniel J. Brat, Gena M. Mastrogianakis, et al. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216):1061–1068, September 2008. doi: 10.1038/nature07385. URL https://doi.org/10.1038/nature07385. (Cited on pages 1 and 39.)

Bjoern H. Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Trans. Med. Imaging.*, 34(10):1993–2024, October 2015. doi: 10.1109/tmi.2014.2377694. (Cited on pages 26 and 39.)

F. Milletari, N. Navab, and S. Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 565–571, Los Alamitos, CA, USA, oct 2016. IEEE Computer Society. doi: 10.1109/3DV.2016.79. URL https://doi.ieeecomputersociety.org/10.1109/3DV.2016.79. (Cited on page 28.)

David Molina, Julián Pérez-Beteta, Alicia Martínez-González, Juan Martino, Carlos Velasquez, Estanislao Arana, and Víctor M. Pérez-García. Lack of robustness of textural measures obtained from 3d brain tumor mris impose a need for standardization. *PLOS ONE*, 12(6):1–14, 06 2017. doi: 10.1371/journal.pone.0178843. URL https://doi.org/10.1371/journal.pone.0178843. (Cited on page 24.)

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. doi: https://doi.org/10.1002/acm2.12795. URL https://aapm.onlinelibrary.wiley.com/doi/abs/10.1002/acm2.12795. (Cited on page 25.)

Andriy Myronenko. 3d MRI brain tumor segmentation using autoencoder regularization. *CoRR*, abs/1810.11654, 2018. URL http://arxiv.org/abs/1810.11654. (Cited on page 29.)

Ingrid Måge, Bjørn-Helge Mevik, and Tormod Næs. Regression models with process variables and parallel blocks of raw material measurements. *Journal of Chemometrics*, 22(8):443–456, 2008. doi: https://doi.org/10.1002/cem.1169. URL https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/abs/10.1002/cem.1169. (Cited on page 31.)

Christophe Nioche, Fanny Orlhac, Sarah Boughdad, Sylvain Reuzé, Jessica Goya-Outi, Charlotte Robert, et al. LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity. *Cancer Research*, 78 (16):4786–4789, 08 2018. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-18-0125. URL https://doi.org/10.1158/0008-5472.CAN-18-0125. (Cited on page 12.)

A.G. Osborn, D.N. Louis, T.Y. Poussaint, L.L. Linscott, and K.L. Salzman. The 2021 world health organization classification of tumors of the central nervous system: What neuroradiologists need to know. *American Journal of Neuroradiology*, 43(7):928–937, June 2022. doi: 10.3174/ajnr.a7462. URL https://doi.org/10.3174/ajnr.a7462. (Cited on page 2.)

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. (Cited on page 56.)

F.R.S. Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901. doi: 10.1080/14786440109462720. (Cited on page 19.)

Jurgen Peerlings, Henry C. Woodruff, Jessica M. Winfield, Abdalla Ibrahim, Bernard E. Van Beers, Arend Heerschap, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Scientific Reports*, 9(1):4800, Mar 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-41344-5. URL https://doi.org/10.1038/s41598-019-41344-5. (Cited on page 25.)

Cathy Philippe. *Analyse intégrée de données de génomique et d'imagerie pour le diagnostic et le suivi du gliome malin chez l'enfant*. PhD thesis, 2014. URL http://www.theses.fr/2014PA112368. Thèse de doctorat dirigée par Frouin, Vincent Physique Paris 11 2014. (Cited on page 3.)

M. G. Poirot, M. W. A. Caan, H. G. Ruhe, A. Bjørnerud, I. Groote, L. Reneman, and H. A. Marquering. Robustness of radiomics to variations in segmentation methods in multimodal brain mri. *Scientific Reports*, 12(1):16712, Oct 2022. ISSN 2045-2322. doi: 10.1038/s41598-022-20703-9. URL https://doi.org/10.1038/s41598-022-20703-9. (Cited on page 25.)

Stephanie Puget, Cathy Philippe, Dorine A. Bax, Bastien Job, Pascale Varlet, Marie-Pierre Junier, et al. Mesenchymal transition and pdgfra amplification/mutation are key distinct oncogenic events in pediatric diffuse intrinsic pontine gliomas. *PLOS ONE*, 7(2):1–14, 02 2012. doi: 10.1371/journal.pone.0030313. URL https://doi.org/10.1371/journal.pone.0030313. (Cited on page 1.)

Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8(1), February 2007. doi: 10.1186/1471-2105-8-35. URL https://doi.org/10.1186/1471-2105-8-35. (Cited on page 5.)

Julian S. Rechberger, Victor M. Lu, Liang Zhang, Erica A. Power, and David J. Daniels. Clinical trials for diffuse intrinsic pontine glioma: the current state of affairs. *Child's Nervous System*, 36 (1):39–46, Jan 2020. ISSN 1433-0350. doi: 10.1007/s00381-019-04363-1. URL https://doi.org/10.1007/s00381-019-04363-1. (Cited on page 2.)

Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, 2016. doi: 10.1109/CVPR.2016.91. (Cited on pages 52 and 123.)

Syed Reza and Khan Iftekharuddin. Multi-class abnormal brain tissue segmentation using texture features. *Proc NCI MICCAI-BRATS*, 2013:38–42, 01 2013. (Cited on page 26.)

Marvin N. Wright Roman Hornung. *blockForest: Block Forests: Random Forests for Blocks of Clinical and Omics Covariate Data*, 2022. URL https://CRAN.R-project.org/package=blockForest. R package version 0.2.5. (Cited on page 94.)

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4. (Cited on pages 28, 52, 56, and 123.)

El Jurdi Rosana, Caroline Petitjean, Paul Honeine, and Fahed Abdallah. BB-UNet: U-net with bounding box prior. *IEEE Journal of Selected Topics in Signal Processing*, 14(6):1189–1198, October 2020. doi: 10.1109/jstsp.2020.3001502. (Cited on pages 29, 52, and 56.)

Scott Ryall, Uri Tabori, and Cynthia Hawkins. Pediatric low-grade glioma in the era of molecular diagnostics. *Acta Neuropathologica Communications*, 8(1):30, Mar 2020. ISSN 2051-5960. doi: 10.1186/s40478-020-00902-z. URL https://doi.org/10.1186/s40478-020-00902-z. (Cited on page 110.)

Mark R. Segal. Microarray gene expression data with linked survival phenotypes: diffuse large-B-cell lymphoma revisited. *Biostatistics*, 7(2):268–285, 11 2005. ISSN 1465-4644. doi: 10.1093/biostatistics/kxj006. URL https://doi.org/10.1093/biostatistics/kxj006. (Cited on page 80.)

Mohsen Shahhosseini, Guiping Hu, and Hieu Pham. Optimizing ensemble weights and hyperparameters of machine learning models for regression problems, 2020. (Cited on page 54.)

Yiyuan She. Sparse regression with exact clustering. *Electronic Journal of Statistics*, 4(none):1055 – 1096, 2010. doi: 10.1214/10-EJS578. URL https://doi.org/10.1214/10-EJS578. (Cited on page 33.)

Russell T. Shinohara, Elizabeth M. Sweeney, Jeff Goldsmith, Navid Shiee, Farrah J. Mateen, Peter A. Calabresi, et al. Statistical normalization techniques for magnetic resonance imaging. *NeuroImage: Clinical*, 6:9–19, 2014. ISSN 2213-1582. doi: https://doi.org/10.1016/j.nicl.2014.08.008. URL https://www.sciencedirect.com/science/article/pii/S221315821400117X. (Cited on page 44.)

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *J. Stat. Softw.*, 39(5):1–13, March 2011a. (Cited on page 18.)

Noah Simon, Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13, 2011b. doi: 10.18637/jss.v039.i05. URL https://www.jstatsoft.org/v39/i05/. (Cited on page 96.)

J.G. Sled, A.P. Zijdenbos, and A.C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, 1998. doi: 10.1109/42.668698. (Cited on page 44.)

Stephen M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002. doi: https://doi.org/10.1002/hbm.10062. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.10062. (Cited on page 44.)

Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005. doi: 10.1073/pnas.0506580102. URL https://www.pnas.org/doi/abs/10.1073/pnas.0506580102. (Cited on page 81.)

Lydia T Tam, Kristen W Yeom, Jason N Wright, Alok Jaju, Alireza Radmanesh, Michelle Han, et al. MRI-based radiomics for prognosis of pediatric diffuse intrinsic pontine glioma: an international study. *Neuro-Oncology Advances*, 3(1), 03 2021. ISSN 2632-2498. doi: 10.1093/noajnl/vdab042. URL https://doi.org/10.1093/noajnl/vdab042. vdab042. (Cited on page 24.)

TCGA. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine*, 372(26):2481–2498, June 2015. doi: 10.1056/nejmoa1402121. URL https://doi.org/10.1056/nejmoa1402121. (Cited on pages 1, 6, and 40.)

Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257–284, March 2011. doi: 10.1007/s11336-011-9206-8. URL https://doi.org/10.1007/s11336-011-9206-8. (Cited on pages 30 and 71.)

Arthur Tenenhaus, Cathy Philippe, Vincent Guillemot, Kim-Anh Le Cao, Jacques Grill, and Vincent Frouin. Variable selection for generalized canonical correlation analysis. *Biostatistics*, 15(3):569–583, 02 2014. ISSN 1465-4644. doi: 10.1093/biostatistics/kxu001. URL https://doi.org/10.1093/biostatistics/kxu001. (Cited on pages 30, 72, 108, and 124.)

Andrew E. Teschendorff, Han Jing, Dirk S. Paul, Joni Virta, and Klaus Nordhausen. Tensorial blind source separation for improved analysis of multi-omic data. *Genome Biology*, 19(1):76, Jun 2018. ISSN 1474-760X. doi: 10.1186/s13059-018-1455-8. URL https://doi.org/10.1186/s13059-018-1455-8. (Cited on page 30.)

Robert Tibshirani. Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, January 1996. ISSN 00359246. doi: 10.1111/j.2517-6161.1996.tb02080.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1996.tb02080.x. (Cited on pages 16 and 32.)

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005. doi: https://doi.org/10.1111/j.1467-9868.2005.00490.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00490.x. (Cited on page 32.)

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. doi: 10.1109/TMI.2010.2046908. (Cited on page 44.)

Hyemin Um, Florent Tixier, Dalton Bermudez, Joseph O Deasy, Robert J Young, and Harini Veeraraghavan. Impact of image preprocessing on the scanner dependence of multi-parametric mri radiomic features and covariate shift in multi-institutional glioblastoma datasets. *Physics in Medicine & Biology*, 64(16):165011, aug 2019. doi: 10.1088/1361-6560/ab2f44. URL https://dx.doi.org/10.1088/1361-6560/ab2f44. (Cited on page 24.)

Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, Nicole Aucoin, Vivek Narayan, et al. Computational Radiomics System to Decode the Radiographic Phenotype. *Cancer Research*, 77(21):e104–e107, 10 2017. ISSN 0008-5472. doi: 10.1158/0008-5472.CAN-17-0339. URL https://doi.org/10.1158/0008-5472.CAN-17-0339. (Cited on pages 12 and 45.)

Guido Van Rossum and Fred L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009. ISBN 1441412697. (Cited on page 45.)

Janita E. van Timmeren, Davide Cester, Stephanie Tanadini-Lang, Hatem Alkadhi, and Bettina Baessler. Radiomics in medical imaging—"how-to" guide and critical reflection. *Insights into Imaging*, 11(1):91, Aug 2020. ISSN 1869-4101. doi: 10.1186/s13244-020-00887-2. URL https://doi.org/10.1186/s13244-020-00887-2. (Cited on pages 24 and 26.)

Fabio Vandin, Eli Upfal, and Benjamin J Raphael. Algorithms for detecting significantly mutated pathways in cancer. *J. Comput. Biol.*, 18(3):507–522, March 2011a. (Cited on page 5.)

Fabio Vandin, Eli Upfal, and Benjamin J. Raphael. Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*, 18(3):507–522, 2011b. doi: 10.1089/cmb.2010.0265. URL https://doi.org/10.1089/cmb.2010.0265. PMID: 21385051. (Cited on page 46.)

Oron Vanunu, Oded Magger, Eytan Ruppin, Tomer Shlomi, and Roded Sharan. Associating genes and protein complexes with disease via network propagation. *PLOS Computational Biology*, 6(1):1–9, 01 2010. doi: 10.1371/journal.pcbi.1000641. URL https://doi.org/10.1371/journal.pcbi.1000641. (Cited on page 46.)

Charles J Vaske, Stephen C Benz, J Zachary Sanborn, Dent Earl, Christopher Szeto, Jingchun Zhu, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, 26(12):i237–45, June 2010. (Cited on page 5.)

Javier E. Villanueva-Meyer, Marc C. Mabray, and Soonmee Cha. Current clinical brain tumor imaging. *Neurosurgery*, 81(3):397–415, May 2017. doi: 10.1093/neuros/nyx103. (Cited on page 51.)

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, et al. SciPy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17(3):261–272, February 2020. doi: 10.1038/s41592-019-0686-2. (Cited on page 54.)

M.W. Wagner, N. Hainc, F. Khalvati, K. Namdar, L. Figueiredo, M. Sheng, et al. Radiomics of pediatric low-grade gliomas: Toward a pretherapeutic differentiation of braf-mutated and braf-fused tumors. *American Journal of Neuroradiology*, 42(4):759–765, 2021. ISSN 0195-6108. doi: 10.3174/ajnr.A6998. URL http://www.ajnr.org/content/42/4/759. (Cited on page 24.)

Katherine E. Warren. Diffuse intrinsic pontine glioma: poised for progress. *Front. Oncol.*, 2, 2012. doi: 10.3389/fonc.2012.00205. (Cited on page 51.)

Takanori Watanabe, Daniel Kessler, Clayton Scott, Michael Angstadt, and Chandra Sripada. Disease prediction based on functional connectomes using a scalable and spatially-informed support vector machine. *NeuroImage*, 96:183–202, August 2014. ISSN 1053-8119. doi: 10.1016/j.neuroimage.2014.03.067. URL https://www.sciencedirect.com/science/article/pii/S1053811914002286. (Cited on page 34.)

D. M. Witten, R. Tibshirani, and T. Hastie. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3):515–534, April 2009. doi: 10.1093/biostatistics/kxp008. URL https://doi.org/10.1093/biostatistics/kxp008. (Cited on page 73.)

Derek Wong, Tae Hoon Lee, Amy Lum, Valerie Lan Tao, and Stephen Yip. Integrated proteomic analysis of low-grade gliomas reveals contributions of 1p-19q co-deletion to oligodendroglioma. *Acta Neuropathologica Communications*, 10(1), May 2022. doi: 10.1186/s40478-022-01372-1. URL https://doi.org/10.1186/s40478-022-01372-1. (Cited on page 1.)

Laura D. Wood, D. Williams Parsons, Sia'n Jones, Jimmy Lin, Tobias Sjöblom, Rebecca J. Leary, et al. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853):1108–1113, November 2007. doi: 10.1126/science.1145720. URL https://doi.org/10.1126/science.1145720. (Cited on page 5.)

Zhuorui Xie, Allison Bailey, Maxim V. Kuleshov, Daniel J. B. Clarke, John E. Evangelista, Sherry L. Jenkins, et al. Gene set knowledge discovery with enrichr. *Current Protocols*, 1(3):e90, 2021. doi: https://doi.org/10.1002/cpz1.90. URL https://currentprotocols.onlinelibrary.wiley.com/doi/abs/10.1002/cpz1.90. (Cited on pages 88 and 94.)

Ji'an Yang and Qian Yang. Identification of core genes and screening of potential targets in glioblastoma multiforme by integrated bioinformatic analysis. *Frontiers in Oncology*, 10, 2021. ISSN 2234-943X. doi: 10.3389/fonc.2020.615976. URL https://www.frontiersin.org/articles/10.3389/fonc.2020.615976. (Cited on page 101.)

Zi Yang and George Michailidis. A non-negative matrix factorization method for detecting modules in heterogeneous omics multi-modal data. *Bioinformatics*, 32(1):1–8, 09 2015. ISSN 1367-4803. doi: 10.1093/bioinformatics/btv544. URL https://doi.org/10.1093/bioinformatics/btv544. (Cited on page 29.)

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68:49–67, 02 2006. doi: 10.1111/j.1467-9868.2005.00532.x. (Cited on page 33.)

Rui Zhang, Hui Luo, Shuai Wang, Zhengxin Chen, Lingyang Hua, Hong-Wei Wang, et al. miR-622 suppresses proliferation, invasion and migration by directly targeting activating transcription factor 2 in glioma cells. *Journal of Neuro-Oncology*, 121(1):63–72, September 2014. doi: 10.1007/s11060-014-1607-y. URL https://doi.org/10.1007/s11060-014-1607-y. (Cited on page 88.)

Shihua Zhang, Chun-Chi Liu, Wenyuan Li, Hui Shen, Peter W. Laird, and Xianghong Jasmine Zhou. Discovery of multi-dimensional modules by integrative analysis of cancer genomic data. *Nucleic Acids Research*, 40(19):9379–9391, 08 2012. ISSN 0305-1048. doi: 10.1093/nar/gks725. URL https://doi.org/10.1093/nar/gks725. (Cited on page 29.)

Wei Zhang, Takayo Ota, Viji Shridhar, Jeremy Chien, Baolin Wu, and Rui Kuang. Network-based survival analysis reveals subnetwork signatures for predicting outcomes of ovarian cancer treatment. *PLoS Comput. Biol.*, 9(3):e1002975, March 2013. (Cited on pages 5 and 76.)

Wei Zhang, Jeremy Chien, Jeongsik Yong, and Rui Kuang. Network-based machine learning and graph theory algorithms for precision oncology. *npj Precision Oncology*, 1(1), August 2017. doi: 10.1038/s41698-017-0029-7. URL https://doi.org/10.1038/s41698-017-0029-7. (Cited on page 5.)

Chenhong Zhou, Shengcong Chen, Changxing Ding, and Dacheng Tao. Learning contextual and attentive information for brain tumor segmentation. In Alessandro Crimi, Spyridon Bakas, Hugo Kuijf, Farahani Keyvan, Mauricio Reyes, and Theo van Walsum, editors, *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, pages 497–507, Cham, 2019. Springer International Publishing. ISBN 978-3-030-11726-9. (Cited on page 28.)

Fangfang Zhu, Jiang Li, Juan Liu, and Wenwen Min. Network-based cancer genomic data integration for pattern discovery. *BMC genomic data*, 22(Suppl 1):54, December 2021. ISSN 2730-6844. doi: 10.1186/s12863-021-01004-y. (Cited on pages 5, 34, and 76.)

Marinka Zitnik and Blaz Zupan. Data fusion by matrix factorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(1):41–53, January 2015. doi: 10.1109/tpami.2014.2343973. URL https://doi.org/10.1109/tpami.2014.2343973. (Cited on page 30.)

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.1467-9868.2005.00503.x. URL https://onlinelibrary.wiley.com/doi/10.1111/j.1467-9868.2005.00503.x. (Cited on pages 16 and 119.)