

École polytechnique de Louvain

Overall Survival Prediction in Glioblastoma Patients Using Multimodal MRI and Radiomic Feature Extraction

Author: **Emilie CARELS**

Supervisor: **Benoît MACQ**

Readers: **John LEE, Manon DAUSORT, Nicolas DELINTE**

Academic year 2025–2026

Master [120] in Biomedical Engineering

Abstract

Glioblastoma (GBM) is the most common and most aggressive primary brain tumour in adults, with a median survival of approximately 15 months. As the most appropriate treatment differs from one patient to another, the ability to predict survival is valuable for tailoring patient management. Magnetic resonance imaging (MRI) is widely recognised as a powerful, non-invasive tool for visualising tumour structures within the brain, and radiomics offers a way to extract quantitative information from these images. This work aims to model overall survival (OS) in GBM patients and to identify the imaging biomarkers that carry the prognostic signal. A cohort of 234 pre-operative MRI examinations from the BraTS2020 dataset was preprocessed, from which 10,203 radiomic features were extracted per patient across three tumour sub-regions, together with the patient’s age at diagnosis. This high-dimensional set was reduced to a consensus of nine interpretable features through a selection procedure combining three complementary methods. Eight survival models were then trained and compared, in cross-validation and on a held-out test set kept strictly isolated throughout, and the one best suited to a clinical context was retained. The Random Survival Forest was retained as the best overall model, reaching a test-set concordance index of 0.671, a mean time-dependent AUC of 0.739 and an integrated Brier score of 0.193, performance on a par with the state of the art for conventional-MRI radiomics. It also separated patients into two significantly distinct risk groups (log-rank $p = 1.43 \times 10^{-2}$, a 192-day difference in restricted mean survival time), with the retained signature concentrating in the tumour core. These results show that a model built solely on conventional multimodal MRI can stratify GBM patients by risk at a level competitive with the literature, while preserving the biological interpretability required for clinical translation.

Keywords: Glioblastoma, Radiomics, Survival Analysis, MRI, Machine Learning

Acknowledgements

I am especially thankful to my supervisor, Professor Benoît Macq for giving me the opportunity to work on this topic, and for his support and his availability throughout the year.

I would also like to thank Manon Dausort, Nicolas Delinte and Simon François for the time they dedicated to me, and for their expertise, advice and feedback. Their help has been invaluable in carrying out this project.

I am thankful to my family, whose support throughout my studies has been beyond compare.

I am thankful to my friends, who stood by me and were there for me through the most difficult moments.

Finally, I wish to thank everyone who encouraged me, supported me and believed in me.

This work was completed with the help of several tools. First, computational resources have been provided by the Consortium des Équipements de Calcul Intensif (CÉCI), funded by the Fonds de la Recherche Scientifique de Belgique (F.R.S.-FNRS) under Grant No. 2.5020.11 and by the Walloon Region. Claude [1] was used for text rephrasing, translation and code debugging; DeepL [2] was also used for translation; and Zotero [3] was used to manage the bibliographic references. All the code used to carry out this work is available on GitHub at https://github.com/emiliec420/overall_survival_prediction_in_glioblastoma_patients.

Contents

Abstract	i
Acknowledgements	ii
List of Figures	v
List of Tables	vii
Acronyms	viii
Introduction	1
1 Theoretical background	3
1.1 Magnetic Resonance Imaging	3
1.1.1 From the hydrogen nucleus to the MRI signal	3
1.1.2 Relaxation, sequences and image weighting	5
1.1.3 MRI modalities	7
1.2 Glioblastoma	9
1.2.1 Disease overview and clinical prognosis	9
1.2.2 Tumour regions and imaging modalities	9
1.3 Radiomics	11
1.3.1 Definition and principle	11
1.3.2 The radiomic workflow	11
1.3.3 Radiomic features	12
1.3.4 Image filters	14
1.4 Feature selection	14
1.4.1 Classification of selection methods	15
1.4.2 Feature selection methods	16
1.5 Survival analysis	17
1.5.1 Principles of survival analysis	17
1.5.2 Functions used in survival analysis	18
1.5.3 Survival Models	21
1.5.4 Evaluation metrics	25
1.6 State of the Art	28
2 Materials and Methods	32
2.1 Materials	32
2.2 Methods	34
2.2.1 Image preprocessing	34
2.2.2 Feature extraction	35
2.2.3 Feature selection	37
2.2.4 Model selection	42
2.2.5 Final evaluation	43
3 Results	47
3.1 Feature selection	47
3.2 Survival modelling	60
3.2.1 Cross validation performances per model	61
3.2.2 Test performance per model	63
3.3 Results of the selected model	68

4	Discussion	72
4.1	Each metric answers a different clinical question	72
4.2	RSF selected for clinical use	73
4.3	Comparison with the literature	75
4.4	The tumour core carries the prognosis	76
4.5	Limitations	78
4.6	Future perspectives	80
	Conclusion	81
A	Appendix	93
A.1	Textures Features	93
A.2	All features extracted from PyRadiomics	96
A.3	Results	99

List of Figures

1.1	Magnetization and precession of protons in a static magnetic field $\vec{B}_0$	4
1.2	Microscopic effects of a 90° RF pulse: energy-state transition and spin rephasing.	5
1.3	Longitudinal (T1) and transverse (T2) relaxation after a 90° RF pulse.	6
1.4	Timing diagram of a spin-echo MRI pulse sequence showing the repetition time TR and the echo time TE.	7
1.5	The four MRI modalities used in this work (T1, T1ce, T2, FLAIR), shown on a GBM patient.	8
1.6	MRI modalities and expert sub-region segmentation in glioblastoma	10
1.7	Radiomics workflow	12
1.8	Right-censoring in survival data	18
1.9	Survival function $S(t)$ and Hazard function $h(t)$	18
1.10	Kaplan-Meier survival curves for two patient subgroups	20
2.1	T1 intensity distributions before and after z-score normalisation	35
2.2	Qualitative comparison of a T1 axial slice before and after z-score normalisation	36
2.3	One-standard-deviation rule: component-wise GBM $\times$ permutation importance.	41
3.1	Co-occurrence matrix of correlated feature pairs decomposed by tumour sub-region	48
3.2	Co-occurrence matrix of correlated feature pairs decomposed by MRI modality	48
3.3	Co-occurrence matrix of correlated feature pairs decomposed by radiomic class.	49
3.4	Co-occurrence matrix of correlated feature pairs decomposed by image type.	50
3.5	Distribution of the univariate Cox ranking scores.	51
3.6	Distribution of the permutation-importance scores.	52
3.7	Concordance between univariate Cox and permutation rankings.	53
3.8	Effect of the 1-SD rule on the number of retained features, by combination.	54
3.9	Selection frequency of each feature, decomposed by method.	55
3.10	Composition of the consensus feature set along four axes.	57
3.11	Joint distribution of the nine consensus features along radiomic class $\times$ tumour sub-region.	58
3.12	Joint distribution of the nine consensus features along MRI modality $\times$ tumour sub-region.	58
3.13	Joint distribution of the nine consensus features along radiomic class $\times$ MRI modality.	59
3.14	Spearman correlation matrix between the nine consensus features.	60
3.15	Per-fold distributions of the three metrics in cross-validation.	62
3.16	Bootstrap distributions of the three metrics on the test set.	64
3.17	Time-dependent AUC of the eight calibrated models on the test set	65
3.18	Time-dependent Brier score of the seven calibrated models on the test set.	66
3.19	Kaplan-Meier survival curves of the predicted high- and low-risk groups for each of the eight candidate models on the test set.	67
3.20	RMST of the predicted risk groups at 24 months, by model.	68
3.21	Time-dependent AUC: single consensus features versus full consensus model.	69
3.22	Time-dependent Brier score: single consensus features versus full consensus model.	70

3.23	Permutation importance of the nine consensus features for the retained RSF.	71
A.1	2D example of GLCM construction	93
A.2	2D example of GLRLM construction	93
A.3	2D example of GLSZM construction	94
A.4	2D example of GLDM construction	94
A.5	2D example of NGTDM construction	95
A.6	Optimal k : univariate Cox method, per model.	101
A.7	Optimal k : permutation method, per model.	102

List of Tables

2.1	Baseline characteristics of the BraTS2020 GBM cohort used in this work . .	32
2.2	Summary table of the number of features extracted.	37
2.3	Description of the BraTS2020 GBM cohort	38
2.4	Parameters of the RSF used for permutation importance.	39
2.5	Parameters of the Boruta algorithm and of the internal RSF used as the estimator.	40
2.6	Hyperparameter grids explored using exhaustive search for each model. . . .	44
3.1	Number of features retained at each stage of the selection pipeline	47
3.2	1-SD procedure summary for each method $\times$ model combination	54
3.3	The nine consensus features	56
3.4	Schoenfeld residuals test of the PH assumption for the final multivariate Cox model	61
3.5	Hyperparameter configurations retained for each model	62
3.6	Cross-validated performance of the eight candidate models on the training set.	63
3.7	Test-set performance of the eight calibrated models.	64
A.1	Radiomic features extracted by PyRadiomics.	96
A.2	Top 30 features selected by the univariate Cox approach	99
A.3	Top 30 features selected by permutation importance.	100

Acronyms

1-SD One Standard Deviation.

ADC Apparent Diffusion Coefficient.

AUC(t) Cumulative/Dynamic AUC.

BraTS Multimodal Brain Tumor Segmentation Challenge.

BraTS2020 Multimodal Brain Tumor Segmentation Challenge 2020.

BS Brier Score.

C-index Concordance Index.

CI Confidence Interval.

CoxPH Cox Proportional Hazards.

CSF Cerebrospinal Fluid.

CT Computed Tomography.

CV Cross-Validation.

ED Peritumoral Oedema.

ET Enhancing Tumour.

FLAIR Fluid Attenuated Inversion Recovery.

GBM Glioblastoma.

GBSM Gradient Boosting Survival Models.

GBSM CompWise Gradient Boosting Survival Models Componentwise.

GLCM Gray Level Co-occurrence Matrix.

GLDM Gray Level Dependence Matrix.

GLRLM Gray Level Run Length Matrix.

GLSZM Gray Level Size Zone Matrix.

GTR Gross Total Resection.

HR Hazard Ratio.

IBS Integrated Brier Score.

IBSI Image Biomarker Standardization Initiative.

IPCW Inverse-probability-of-censoring weighting.

IQR Interquartile Range.

KM Kaplan-Meier.

KPS Karnofsky Performance Score.

MRI Magnetic Resonance Imaging.

NA Nelson-Aalen.

NCR Necrotic Core.

NET Non-enhancing Tumour.

NGTDM Neighboring Gray Tone Difference Matrix.

NMR Nuclear Magnetic Resonance.

OS Overall Survival.

PH Proportional Hazards.

RBF Radial Basis Function.

RF Radio Frequency.

RMST Restricted Mean Survival Time.

ROC Receiver Operating Characteristic.

RSF Random Survival Forest.

SSVM Survival Support Vector Machine.

T1 Time 1.

T1ce T1 contrast-enhanced.

T2 Time 2.

TE Echo Time.

TR Repetition Time.

VOI Volumes of Interest.

Introduction

Glioblastoma (GBM) is the most common and the most aggressive primary brain tumour in adults [4, 5]. Even with the current standard of care, maximal safe surgical resection followed by concurrent radiotherapy and temozolomide-based chemotherapy, its prognosis remains poor: median overall survival is approximately 15 months from diagnosis, and no curative treatment exists [4]. The disease predominantly affects older adults, with a median age at diagnosis of around 65 years and an incidence that rises steeply with age [4, 6]. Crucially, patients do not all respond to, or tolerate, aggressive treatment in the same way, so clinical management is individualised rather than uniform [7]. In practice, however, these decisions are too often anchored on chronological age alone, an imperfect predictor: in older patients in particular, frailty and functional status carry prognostic information that age does not [8]. There is therefore a clinical incentive to move towards individualised prognostic assessment, in which an estimate of a given patient’s expected survival can inform a personalised risk–benefit analysis [7].

Magnetic resonance imaging (MRI) is the gold standard for the diagnosis and monitoring of GBM [4], and an experienced radiologist can often recognise the tumour from the images alone. Yet in routine practice this reading is largely qualitative: it relies on the subjective experience of the practitioner, or on coarse quantitative metrics such as the diameter of a lesion [9]. As a result, only a small fraction of the quantitative information contained in the images is actually exploited, and the prognostic judgements drawn from them are inherently subjective and difficult to reproduce or standardise across centers [9]. This is a missed opportunity, because a reliable, individualised survival estimate could be valuable on several clinical fronts [7]. It could help to calibrate the frequency of follow-up imaging; it could support the stratification of patients for clinical trials, for instance by directing those with a particularly poor expected prognosis towards experimental or last-resort protocols; and it could inform prognostic counselling, allowing clinicians to give patients a quantified sense of their outlook, always with due regard for the fact that not every patient wishes to know. What is missing, in short, is an automated, quantitative and interpretable way of turning a routinely acquired MRI examination into a prognostic estimate.

This thesis sets out to build one, using radiomics: the automated extraction of large numbers of quantitative descriptors from medical images, which converts those images into high-dimensional, analysable data [9, 10]. Working with the publicly available BraTS2020 [11, 12, 13, 14, 15] cohort of 234 GBM patients for whom overall survival is known, we extract 10,203 hand-crafted radiomic features per patient from the four conventional MRI sequences (T1, T1ce, T2 and FLAIR) that are available in virtually every clinical protocol, computed over three distinct tumour sub-regions, and we add the patient’s age at diagnosis as the sole clinical covariate. Two methodological choices are made deliberately and in the service of clinical usefulness: hand-crafted features are preferred over deep-learning representations, and feature selection over dimensionality reduction, so that every descriptor retained in the final model keeps a direct and traceable biological meaning. A multi-stage, multi-method selection procedure then reduces this large pool to a compact signature of nine variables. Eight survival models are trained and compared under a common protocol, with the explicit

aim of identifying the model best suited to clinical use rather than the one scoring highest on a single performance metric, and all are evaluated on a held-out test set kept strictly isolated until this final assessment, from which the most suitable model is retained. The work thus pursues two coupled objectives: to build and evaluate a survival model for GBM patients, and to identify interpretable imaging biomarkers associated with their prognosis.

The remainder of this manuscript is organised as follows. Chapter 1 lays the theoretical groundwork, the principles of MRI and its modalities, the biology and clinical prognosis of glioblastoma, the radiomics paradigm, the feature-selection and survival-analysis frameworks on which the methodology rests, and a review of the state of the art in GBM survival prediction. Chapter 2 describes the materials and methods, presenting the BraTS2020 dataset and detailing the full pipeline, from image preprocessing through feature extraction, feature selection and model selection to final evaluation. Chapter 3 reports the results, covering the redundancy structure of the radiomic features, the comparison of the eight candidate models in cross-validation and on the test set, and a closer analysis of the best-performing model. Chapter 4 discusses these findings: how the different evaluation metrics should be read through a clinical lens, the identification of the most suitable model, a comparison with the literature, the link between the selected features and glioblastoma anatomy, and the limitations and future perspectives of the work.

Chapter 1

Theoretical background

1.1 Magnetic Resonance Imaging

Magnetic Resonance Imaging (MRI) is a non-invasive medical imaging technique [16, 17]. It is based on the principles of Nuclear Magnetic Resonance (NMR), discovered by Purcell and Bloch in the 1940s, whose application to medical imaging has profoundly transformed clinical diagnosis [18, 19].

The general principle of MRI can be summarised as follows: it exploits the magnetic properties of hydrogen nuclei present in the tissues of the human body. The detected signal has a complex relationship with the physical and chemical properties of the material from which it originates, allowing for the generation of a wide variety of image contrasts from the same tissue [17].

It is precisely this capability that distinguishes MRI from other medical imaging modalities: while Computed Tomography (CT)¹ essentially provides only a single type of contrast based on the density of tissues, MRI can selectively image several distinct tissue characteristics (such as proton density, molecular environment, water mobility, etc.) from a single scan [20]. A pathology that would remain invisible on a given type of image can thus be detected based on its effect on another property [20]. This flexibility makes MRI the gold standard for brain imaging, particularly for the diagnosis and monitoring of tumours.

1.1.1 From the hydrogen nucleus to the MRI signal

MRI exploits a fundamental quantum property of atomic nuclei: spin, which is an intrinsic quantum property of subatomic particles. It can be pictured as a rotation of the particle around its own axis [21, 22]. Any charged particle with spin behaves like a tiny magnet and carries a magnetic moment $\vec{\mu}$ [21]. Only nuclei with an odd number of protons or neutrons have a non-zero total spin, and are therefore suitable for imaging [18]. Among these, the hydrogen nucleus, which consists of a single proton, is by far the most used for medical imaging [21].

Alignment and precession in a static magnetic field In the absence of an external magnetic field, the magnetic moments $\vec{\mu}$ of the protons in a tissue sample are randomly oriented, and their vector sum is zero ($\sum \vec{\mu} = 0$) [21] (Figure 1.1a, left). When the patient is placed in a strong static field $\vec{B}_0$, conventionally oriented along the z -axis, the protons preferentially align with this field, with a slight excess in the parallel (low-energy) orientation over the antiparallel (high-energy) one. This imbalance gives rise to a non-zero macroscopic

¹CT is an imaging technique that uses X-rays to obtain 2D or 3D images of the body.

magnetization vector $\vec{M}$, aligned with $\vec{B}_0$ [16, 21] (Figure 1.1a, right). Rather than being strictly aligned with $\vec{B}_0$, the protons describe a precession motion around it (Figure 1.1b). The angular frequency of this precession, known as the Larmor frequency, is given by

$$\omega_0 = \gamma B_0, \quad (1.1)$$

where γ is the gyromagnetic ratio, a constant specific to each nucleus [16]. At equilibrium, $\vec{M}$ is parallel to $\vec{B}_0$ and its magnitude is negligible compared to $\vec{B}_0$ itself, so no signal can be measured. To extract a usable signal, the equilibrium must be perturbed.

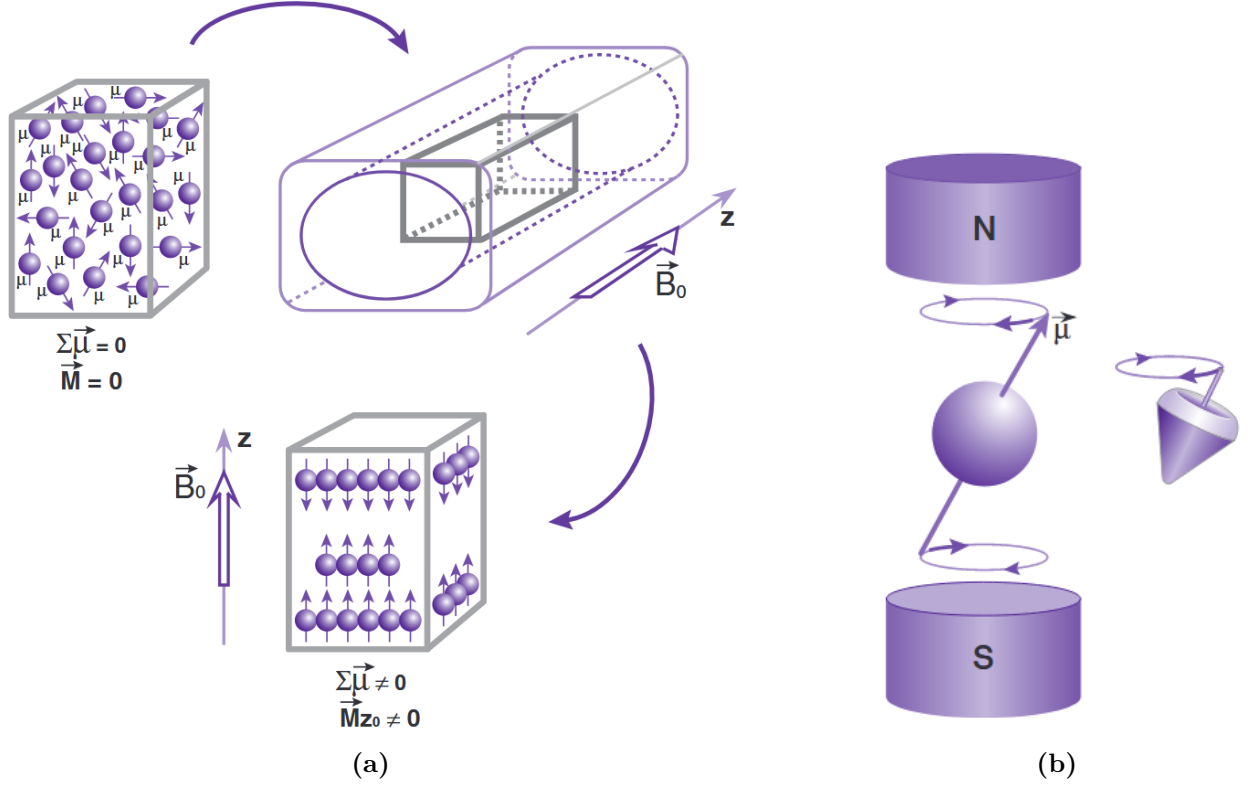


Figure 1.1: (a) In the absence of an external magnetic field (left), the magnetic moments $\vec{\mu}$ of the protons in a tissue sample are randomly oriented, so that their vector sum is zero and no macroscopic magnetization is produced ($\sum \vec{\mu} = 0$, $\vec{M} = 0$). When the sample is placed in a strong static field $\vec{B}_0$ aligned with the z -axis (top right), the protons preferentially align with the field, with a slight excess in the parallel (low-energy) orientation over the antiparallel (high-energy) one. This imbalance gives rise to a non-zero macroscopic magnetization vector $\vec{M}_{z0}$ aligned with $\vec{B}_0$ (bottom right). (b) When placed in an external magnetic field $\vec{B}_0$, the magnetic moment $\vec{\mu}$ of a proton does not align strictly with the field but instead describes a precession motion around it, analogous to the wobble of a spinning top in a gravitational field. Adapted from [21].

RF excitation and resonance The perturbation is delivered by a short electromagnetic pulse in the Radio Frequency (RF) range, tuned to the Larmor frequency of the protons. Because the pulse frequency matches the natural precession frequency, energy is transferred efficiently between the wave and the proton spin system: this is the magnetic resonance phenomenon that gives MRI its name [16]. Under the effect of an RF pulse, the magnetization vector $\vec{M}$ is tilted away from the z -axis. The angle by which $\vec{M}$ is rotated is called the flip angle, and depends on the amplitude and duration of the pulse [21]. Two values play a central role in this work:

- A 90° pulse tips $\vec{M}$ entirely into the transverse plane xy . After such a pulse, the longitudinal component vanishes ($M_z = 0$) and the transverse component M_{xy} is maximal. This is the standard excitation used in conventional T1- and T2-weighted sequences [21].
- A 180° pulse flips $\vec{M}$ to the direction opposite to $\vec{B}_0$. This inversion pulse is used in specialized sequences such as FLAIR (Section 1.1.3), where it is exploited to suppress the signal of a chosen tissue [21].

At the microscopic scale, a 90° pulse simultaneously causes some protons to transition from the low-energy to the high-energy state and brings the precessing protons in phase with one another, which together produce the macroscopic tilt of $\vec{M}$ into the transverse plane (Figure 1.2) [16, 21].

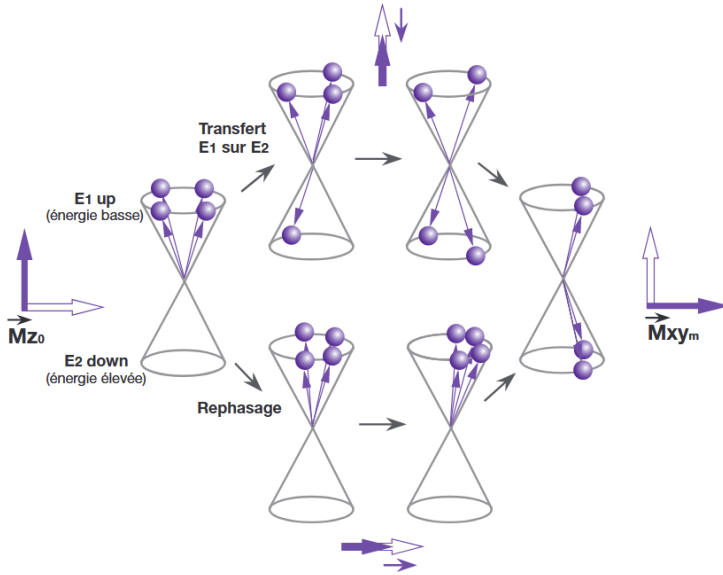


Figure 1.2: The two simultaneous microscopic effects of a 90° RF pulse that bring the magnetization from the longitudinal axis ($\vec{M}_{z0}$, left) to the transverse plane ($\vec{M}_{xy}$, right). Top row: some protons absorb energy from the RF wave and transition from the low-energy state E_1 (parallel to $\vec{B}_0$) to the high-energy state E_2 (antiparallel), cancelling the longitudinal magnetization. Bottom row: the precessing protons become in phase with one another (rephasing), producing a non-zero transverse component. Adapted from [21].

Signal detection It is the rotating transverse component M_{xy} that is actually measured: as it precesses around $\vec{B}_0$ at the Larmor frequency, it induces an alternating current in a receiving coil placed around the patient, and this current constitutes the MRI signal [16]. Once the RF pulse is switched off, the system gradually returns to its equilibrium configuration. This return to equilibrium is called relaxation and is described in the next section 1.1.2. Crucially, its duration depends on the local tissue environment, which is ultimately what gives rise to the contrast seen on MRI images.

1.1.2 Relaxation, sequences and image weighting

When the RF pulse is switched off, the magnetization vector $\vec{M}$ gradually returns to its equilibrium configuration (aligned with $\vec{B}_0$). This return to equilibrium, called relaxation, results from two independent mechanisms that occur simultaneously and are characterized by two distinct time constants, Time 1 (T1) and Time 2 (T2) [23].

Longitudinal relaxation (T1) T1 relaxation describes the gradual recovery of the longitudinal component M_z of the magnetization, as protons return from the high-energy state (antiparallel to $\vec{B}_0$) back to the low-energy state (parallel to $\vec{B}_0$) (Figure 1.3a). The longitudinal component M_z itself induces no signal, T1 effects are observed only indirectly, through the amount of longitudinal magnetization available to be tipped into the transverse plane at the next excitation. T1 relaxation involves an energy transfer from the excited protons to their molecular environment, called the lattice, in the form of heat, which is why it is also referred to as spin-lattice relaxation [18, 23]. The time constant T1 reflects how efficiently the local molecular environment can absorb energy from the excited protons, and depends strongly on tissue composition [21].

Transverse relaxation (T2) T2 relaxation describes the decay of the transverse component M_{xy} . Immediately after a 90° pulse, the protons precess in phase and M_{xy} is maximal. They then gradually dephase due to small local variations in the magnetic field caused by interactions between neighbouring spins, which is why this mechanism is also called spin-spin relaxation (Figure 1.3b) [16, 21]. Unlike T1 relaxation, T2 relaxation does not involve any energy exchange with the environment: it is purely a loss of phase coherence [18].

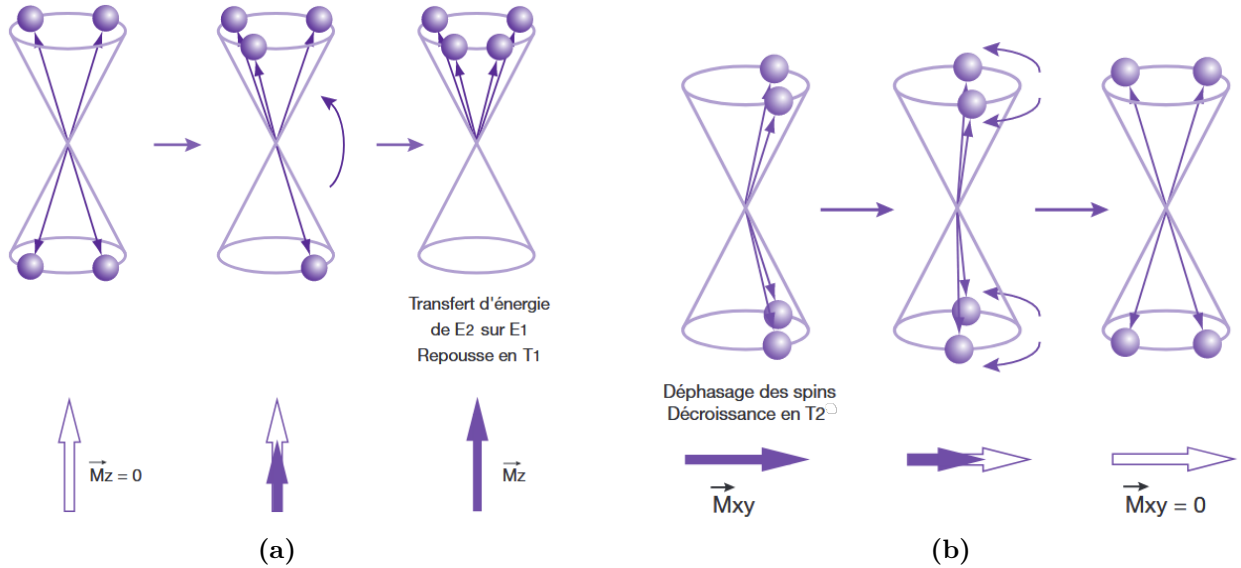


Figure 1.3: (a) Longitudinal (T1) relaxation after a 90° RF pulse. Immediately after the pulse (left), $\vec{M}_z = 0$ and the protons are equally distributed between the low-energy state E_1 (parallel to $\vec{B}_0$) and the high-energy state E_2 (antiparallel). Over time (middle, right), protons in E_2 progressively release their excess energy to the surrounding molecular environment and return to E_1 , so that $\vec{M}_z$ gradually recovers towards its equilibrium value. (b) Transverse (T2) relaxation after a 90° RF pulse. Immediately after the RF pulse (left), all precessing protons are in phase, producing a maximal transverse magnetization $\vec{M}_{xy}$. Over time (middle, right), small local field variations induced by spin-spin interactions cause a progressive loss of phase coherence (dephasing), and the transverse magnetization decays to zero. Adapted from [21].

The origin of tissue contrast T1 and T2 both depend on the molecular composition of the tissue, particularly on the proportions of free water, bound water, fat, and macromolecules. Pure water, for example, has a long T1 because its small molecules tumble too rapidly to exchange energy efficiently with protons; fat molecules, which are larger and slower, have

molecular motion frequencies closer to the Larmor frequency and therefore relax much faster [18]. These tissue-specific differences in T1 and T2 are the fundamental source of the soft-tissue contrast in MRI [23].

MRI sequences, TR and TE To produce an image, a single RF pulse is not enough: the basic scheme is repeated many times to fill the image, with the signal measured shortly after each excitation. The succession of RF pulses and measurements is called a pulse sequence and is governed by two timing parameters that together determine which relaxation mechanism dominates the contrast in the final image [16, 24] (Figure 1.4):

The Repetition Time (TR) is the interval between two successive 90° RF pulses. A short TR does not leave enough time for tissues with long T1 to fully recover their longitudinal magnetization between two pulses, while tissues with short T1 recover almost completely. The former therefore produce a weaker signal and appear dark, while the latter appear hyperintense. The resulting image is said to be T1-weighted (Section 1.1.3) [24]. The Echo Time (TE) is the delay between the RF pulse and the moment at which the signal is measured. A long TE lets the transverse magnetization decay before measurement: tissues with short T2 have already lost most of their signal and appear dark, while tissues with long T2 retain a detectable signal and appear hyperintense. The resulting image is said to be T2-weighted (Section 1.1.3) [18, 24].

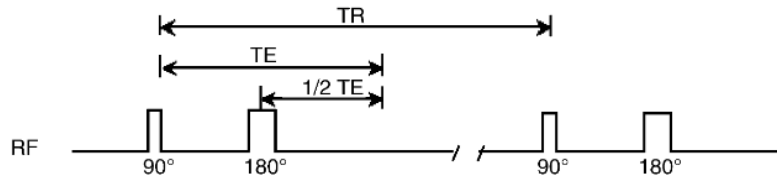


Figure 1.4: Timing diagram of a spin-echo MRI sequence. Each repetition starts with a 90° excitation pulse that tilts the magnetization into the transverse plane; after a delay of $TE/2$, a 180° refocusing pulse compensates for static field inhomogeneities, and the signal is measured at the echo time TE. The whole pattern is repeated every TR seconds. TR controls the T1 weighting of the image and TE its T2 weighting. Adapted from [18]

It is precisely this flexibility that constitutes the major diagnostic advantage of MRI: from a single anatomical examination, several images with different contrasts can be reconstructed, each highlighting different tissue properties [17, 20]. A pathology that is barely visible on one weighting may be conspicuous on another [20]. This property justifies the combined use of several MRI sequences in brain imaging, as we will see in the next section 1.1.3.

1.1.3 MRI modalities

The ability to tune image contrast via TR and TE, combined with the use of contrast agents and specialised pulse schemes, gives rise to a wide variety of complementary MRI sequences. In brain imaging, and particularly in the diagnosis and follow-up of tumours such as the glioblastoma, four sequences are routinely acquired: T1, T2, T1 with gadolinium injection (T1ce), and FLAIR [25]. Figure 1.5 shows an axial slice of the same patient in each of the four modalities, illustrating the complementary information they provide.

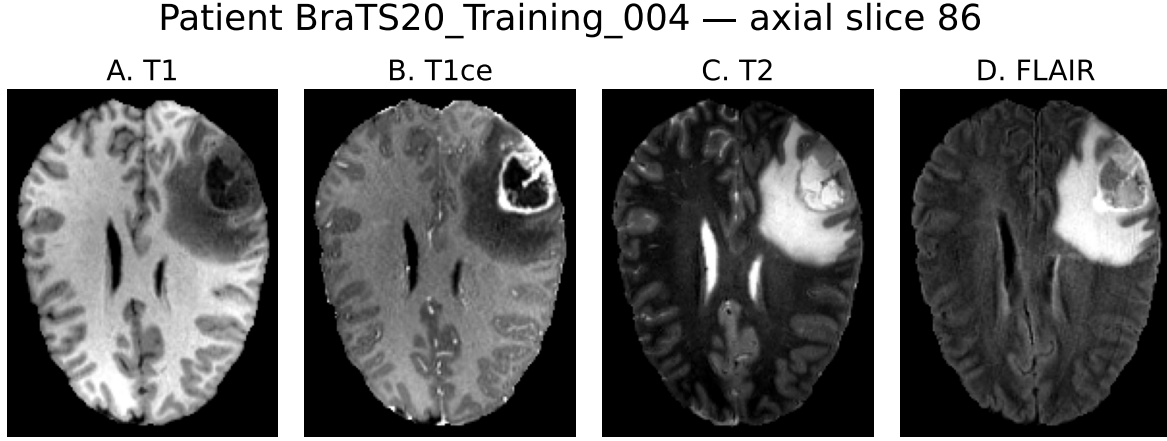


Figure 1.5: The four MRI modalities used in this work, shown on the same axial slice of a GBM patient from the BraTS2020 dataset [11, 12, 13] (patient 004, slice 86). (A) Native T1-weighted image. (B) T1ce (T1-weighted image acquired after intravenous injection of a gadolinium-based contrast agent). (C) T2-weighted image. (D) FLAIR.

T1-weighted A T1-weighted image is acquired with a short TR and a short TE, so that the contrast is dominated by differences in T1 between tissues [24]. On these images, fat appears hyperintense (bright) while fluids, such as the Cerebrospinal Fluid (CSF) in the cerebral ventricles, appear hypointense (dark) [24]. T1-weighted imaging provides an excellent level of anatomical detail and, in particular, allows for clear differentiation between the brain’s white matter and grey matter [25]. It serves as the gold standard for visualizing the normal structure of the brain (Figure 1.5 A).

T1 with gadolinium contrast (T1ce) The T1 contrast-enhanced (T1ce) sequence is a T1-weighted acquisition performed after intravenous injection of a gadolinium-based contrast agent [23]. Gadolinium is a paramagnetic substance that interacts with surrounding protons and substantially shortens their T1. Tissues in which it accumulates therefore appear markedly hyperintense compared to a native T1 image [23] (Figure 1.5 B).

T2-weighted A T2-weighted image is obtained with a long TR and a long TE, so that the contrast is dominated by differences in T2 between tissues [24]. In contrast to T1-weighted imaging, fluids now appear hyperintense, including the CSF in the ventricles and subarachnoid spaces [24]. Any pathological process that locally increases tissue water content, such as oedema, inflammation, or tumour infiltration, also appears hyperintense [19, 25]. This makes T2-weighted imaging particularly useful for detecting pathology associated with changes in tissue water composition (Figure 1.5 C).

Fluid Attenuated Inversion Recovery (FLAIR) FLAIR is a variant of T2-weighted imaging in which the CSF signal is specifically suppressed [25]. The result is a strongly T2-weighted image in which free fluids appear dark instead of bright. On a standard T2 image, peritumoural oedema and CSF are both hyperintense, which makes it difficult to separate the two, particularly where the oedema abuts the ventricles or subarachnoid spaces. By suppressing the CSF signal, FLAIR allows for clear delineation of peritumoural oedema and tumour infiltration, especially in periventricular regions [18, 25] (Figure 1.5 D).

1.2 Glioblastoma

1.2.1 Disease overview and clinical prognosis

Glioblastoma (GBM) is the most common and most aggressive primary brain tumour in adults. It belongs to the family of diffuse gliomas, alongside lower-grade astrocytomas and oligodendrogliomas, and arises from glial cells of the central nervous system [4, 5]. The annual incidence is 3 to 5 cases per 100,000 people and increases markedly with age, reaching 15.2 per 100,000 in the 75–84 age group, the median age at diagnosis is around 65 years [4, 6].

When GBM is suspected, the gold standard test is brain MRI. A CT scan with intravenous contrast may be performed when MRI is not available. However, MRI is preferred in most cases because its sensitivity for detecting changes specific to GBM is higher than that of CT, and it does not require radiation exposure [4, 6]. With experience, healthcare professionals are able to diagnose GBM based on MRI images alone. Confirmation can, however, be obtained from a tissue sample taken during surgical resection or, in elderly patients in poor general condition, via biopsy [4].

The standard of care combines maximal safe surgical resection followed by concurrent radiotherapy and temozolomide-based chemotherapy [5, 26]. Despite this multimodal treatment, the prognosis remains poor: median overall survival is approximately 15 months from diagnosis, and there is currently no curative therapy [4]. In practice, treatment is individualised rather than uniform, and such decisions are often based on chronological age [8]. Age taken on its own, however, is an imperfect criterion: in elderly patients, frailty and functional status predict survival independently of chronological age and are increasingly regarded as more informative for therapeutic decision-making than age alone [8]. This has motivated a shift towards individualised prognostic assessment, in which models that estimate an individual patient’s expected survival, integrating age together with clinical, imaging and molecular factors, are advocated as a way to tailor clinical management through a personalised risk–benefit analysis [7].

1.2.2 Tumour regions and imaging modalities

A defining characteristic of GBM is its high degree of intratumoural heterogeneity: the tumour is not a uniform mass but a collection of biologically distinct sub-regions [12]. This heterogeneity is itself prognostically meaningful, as it correlates with tumour aggressiveness and treatment resistance [27], and, crucially, it is directly visible on standard multiparametric MRI. Each sub-region produces a specific signature across the four modalities introduced in Section 1.1.3:

- **Necrotic Core (NCR)** results from local hypoxia: when tumour growth exceeds the capacity of the available blood supply, cells die due to a lack of oxygen and nutrients. It appears hypointense on T1ce MRI compared to native T1 [12] (Figure 1.6 A and B, yellow).
- **Non-enhancing Tumour (NET)** corresponds to the solid portion of the tumour whose vascularisation is insufficient to allow for significant gadolinium leakage. Similar

to necrosis, this area also appears hypointense on T1-weighted images, making the two regions difficult to distinguish visually [12] (Figure 1.6 A and B, yellow).

- **Enhancing Tumour (ET)** corresponds to active, highly vascularised tumour cells. The enhancement results from a breach in the blood-brain barrier caused by tumour neovascularisation, which allows the contrast agent (gadolinium) to leak and accumulate locally [28]. This area appears hyperintense on T1ce compared to native T1 [12] (Figure 1.6 A and B, red).
- **Peritumoral Oedema (ED)** is the region surrounding the tumour core, containing both vasogenic oedema and infiltrating tumour cells, two components indistinguishable on standard MRI [6, 28]. It appears hyperintense on T2 and, even more clearly, on FLAIR, where CSF signal suppression isolates the oedema from neighbouring fluid-filled structures [25] (Figure 1.6 C and D, green).

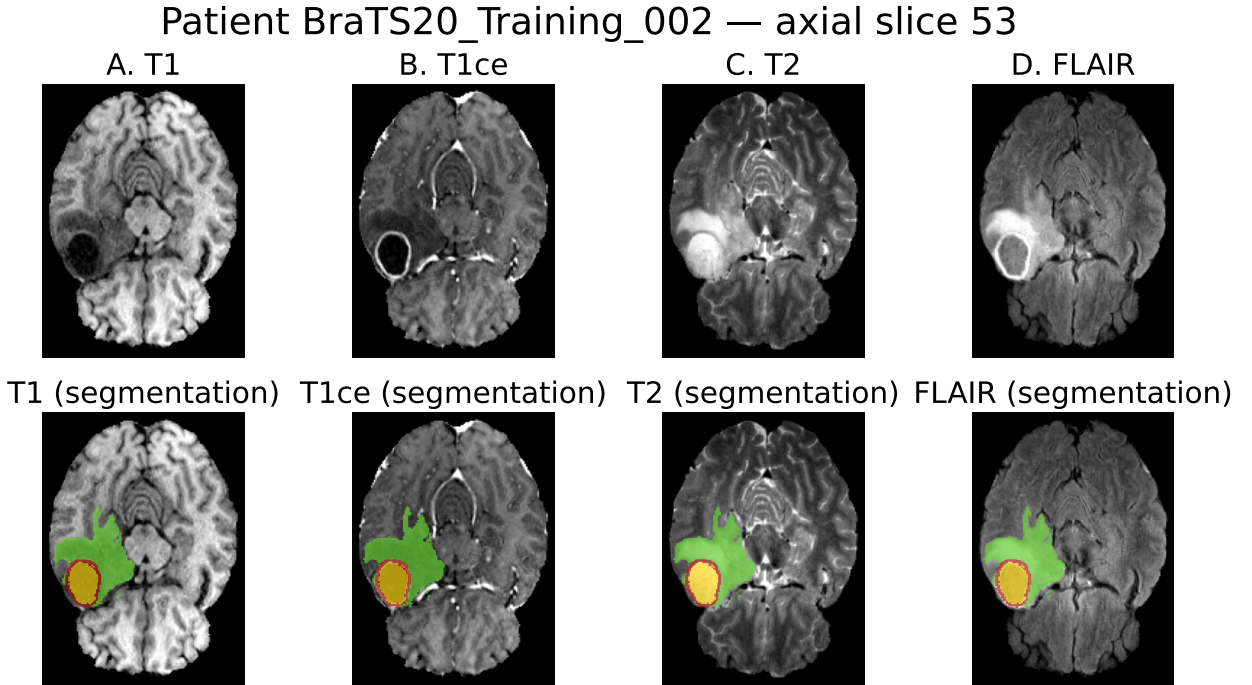


Figure 1.6: The four MRI modalities (T1, T1ce, T2, FLAIR) acquired for a representative patient (top row), and the corresponding expert sub-region segmentation overlaid on each modality (bottom row). The necrotic and non-enhancing tumour core appears in yellow, the enhancing tumour in red, and the peritumoural oedema in green.

The precise annotation protocol used to delineate four sub-regions in the data analysed in this thesis is detailed in Section 2.1.

1.3 Radiomics

1.3.1 Definition and principle

In clinical practice, the interpretation of radiological images generally relies on a qualitative assessment, often based on the practitioner’s subjective experience, or on simple metrics such as the diameter of a lesion. As a result, only a fraction of the quantitative information contained in medical images is utilised in clinical workflows [9].

Radiomics is a quantitative approach to medical imaging that automatically extracts a large number of features from medical images, thereby converting them into high-dimensional, mineable data [9, 10]. The fundamental assumption of radiomics is that medical images contain disease-specific information that is imperceptible to the human eye [29]. Through the mathematical extraction of the spatial distribution of signal intensities and voxel relationships, radiomics quantifies textural information using methods derived from the field of artificial intelligence [29].

Radiomic features can be categorised into two main types: hand-crafted features, obtained using traditional algorithms (intensity, shape, texture, wavelets), and features generated by deep learning algorithms [30].

GBM is a particularly compelling target for radiomic analysis. As described in Section 1.2, it is a spatially heterogeneous tumour, exhibiting on a single MRI examination several distinct sub-regions that coexist within the same lesion [12].

In this context, the limitations of conventional molecular characterization become apparent: a biopsy samples only a small fraction of the tumour and may not capture the regions driving the patient’s prognosis [31]. Multiparametric MRI, in contrast, is acquired as part of routine pre-operative practice and covers the entire tumour volume non-invasively, which makes it an ideal substrate for the quantitative analysis of intratumoural heterogeneity, and ultimately for the survival modeling pursued in this work.

1.3.2 The radiomic workflow

The radiomics pipeline consists of five distinct stages, each presenting specific challenges [30]: image acquisition and preprocessing, segmentation, feature extraction, feature selection, and finally model building and evaluation. This workflow is illustrated in Figure 1.7

Image acquisition and preprocessing Many intrinsic factors related to image acquisition can influence the resulting radiomic signatures, including the equipment used (scanner brand and model), acquisition parameters (magnetic field strength and sequences for MRI, administration of contrast agent), and reconstruction parameters (filters, voxel size) [9]. One of the main challenges at this stage is the high variability in acquisition protocols across different medical centers, which can compromise the reproducibility of quantitative features and, consequently, the external validity of radiomic models [30]. After acquisition, several preprocessing steps are required to harmonise the images, like voxel spacing, normalization of gray-level intensities or discretization [29].

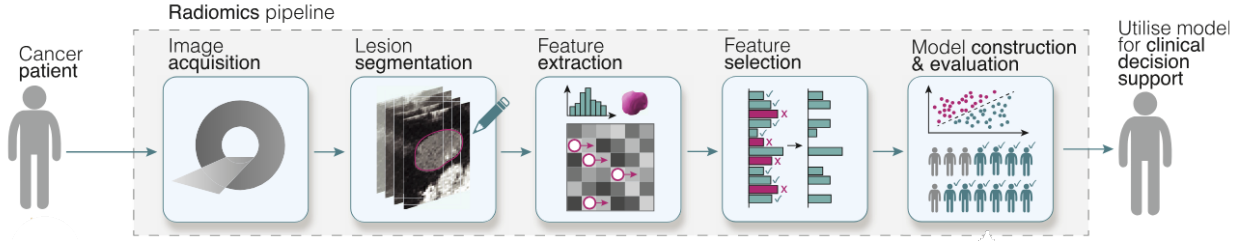


Figure 1.7: The five sequential steps of the radiomics pipeline applied to a cancer patient. Multimodal medical images are first acquired, the lesion is segmented to define the region of interest, quantitative features are extracted from this region, the most informative features are selected to reduce dimensionality, and a predictive model is finally built and evaluated to support clinical decision-making. Adapted from [9].

Segmentation Before any analysis can be performed, the images must be segmented; that is, the regions from which features will be extracted must be delimited [9]. Radiomic features are significantly influenced not only by image acquisition but also by variability in Volumes of Interest (VOI) boundary definition. Accurate segmentation is therefore a pivotal step in radiomic studies because it significantly affects derived characteristics [30]. AI-based automated segmentation is considered to partially mitigate the problem of inter-reader variability; however, it is nevertheless recommended that these automatic segmentations be reviewed by a specialised reader if necessary [9].

Extraction of radiomic features Feature extraction involves calculating quantitative features from the grayscale values within the VOI. It is recommended to follow the guidelines of the Image Biomarker Standardization Initiative (IBSI), which provide a consensus for standardised feature calculations [29]. More information about the various types of radiomic features will be developed in Section 1.3.3.

Feature selection Dimensionality reduction is a process that identifies the least redundant features, thereby contributing to the construction of a generalizable radiomic model. More information on this part will be given in Section 1.4.

Model development and evaluation Once the features have been selected, they can be used in various diagnostic, prognostic, or predictive models [30]. This is the core of this work and will be developed in Section 1.5.

1.3.3 Radiomic features

Features are conventionally grouped into three families: first-order, shape, and texture features [27, 32].

First-order features First-order features describe the global distribution of voxel intensities inside the VOI, treating each voxel independently and ignoring any spatial relationship between them. They are derived from the intensity histogram and include simple descriptors such as the mean, median, variance, skewness, kurtosis, energy, and entropy. As such,

they characterise the amount and dispersion of signal within the lesion, but not its spatial organisation [27, 32].

Shape features Shape features describe the geometry of the VOI itself, independently of voxel intensities. They are computed solely from the segmentation mask and include size descriptors (volume, surface area, maximum 3D diameter) and morphology descriptors (sphericity, compactness, surface-to-volume ratio, elongation, flatness). In oncology, shape features are clinically meaningful because tumour shape regularity is often associated with biological behaviour: spiculated or irregular tumours tend to be more infiltrative and aggressive [27, 32].

Texture features Texture features describe the spatial distribution of intensities, that is, how voxels of similar or different gray levels are organised relative to one another. Unlike first-order statistics, they capture inter-voxel relationships and therefore quantify intratumoural heterogeneity, which is one of the key biological motivations for radiomics. Five complementary matrices are commonly used, each encoding a different "view" of texture.

- **Gray Level Co-occurrence Matrix (GLCM).** The GLCM $P(i, j | \delta, \theta)$ counts the number of times a voxel pair of gray levels (i, j) occurs in the image at a given distance δ and direction θ [32]. For an image quantised to N_g gray levels, the GLCM is therefore an $N_g \times N_g$ matrix. An example is given in Figure A.1.
- **Gray Level Run Length Matrix (GLRLM).** The GLRLM $P(i, j | \theta)$ counts the number of runs of length j of voxels having gray level i , along a given direction θ [32]. A run is a sequence of consecutive voxels sharing the same gray level. The GLRLM therefore captures the coarseness of the texture along a direction: long runs of identical gray levels indicate smooth, homogeneous regions, whereas a predominance of short runs indicates a fine, busy texture. An example is given in Figure A.2.
- **Gray Level Size Zone Matrix (GLSZM).** The GLSZM $P(i, j)$ counts the number of zones of size j voxels with gray level i , where a zone is defined as a set of connected voxels (26-connectivity in 3D) sharing the same gray level [32]. Unlike GLCM and GLRLM, the GLSZM is by construction direction-independent and captures 3D regional homogeneity rather than directional runs. Large, uniform zones translate to a few large-area-emphasis entries, while highly heterogeneous tissue produces many small zones. An example is given in Figure A.3.
- **Gray Level Dependence Matrix (GLDM).** The GLDM $P(i, j)$ counts the number of voxels of gray level i that have exactly j dependent voxels in their neighbourhood [32]. A voxel k in the neighbourhood of voxel c is considered dependent on c if $|I(k) - I(c)| \leq \alpha$, where α is a tolerance threshold. The GLDM thereby characterises how strongly each voxel is "locally supported" by its neighbours: high dependence indicates locally homogeneous regions, low dependence indicates rapid intensity changes. An example is given in Figure A.4.
- **Neighboring Gray Tone Difference Matrix (NGTDM).** The NGTDM is a one-dimensional matrix that sums, for each gray level i , the absolute differences between voxels of gray level i and the mean gray level of their neighbourhood (typically a $3 \times 3 \times 3$

window) [32]. It encodes the perceptual aspects of texture, in particular coarseness, contrast, busyness, complexity, and strength. NGTDM features are therefore closer to how a human observer would intuitively describe a texture. An example is given in Figure A.5.

1.3.4 Image filters

Computing radiomic features directly on the original image captures texture at the native spatial scale of the acquisition. However, intratumoural heterogeneity manifests at multiple scales and frequencies, and certain biologically relevant patterns may not be optimally captured by a single-scale analysis. Image filters are therefore commonly applied before feature extraction to enhance specific spatial-frequency components of the image, so that the same descriptors can be evaluated at several scales [33].

Wavelet decomposition The discrete wavelet transform decomposes an image into components at different scales and orientations through the iterated application of complementary filters: a low-pass filter L , which retains coarse, low-frequency content, and a high-pass filter H , which retains fine, high-frequency content [32]. For a 3D volume, the decomposition is applied independently along each of the three spatial axes, producing $2^3 = 8$ sub-bands corresponding to all combinations of L and H in the three directions: LLL, LLH, LHL, LHH, HLL, HLH, HHL, and HHH [32]. Each sub-band is itself a 3D image of the same dimensionality as the original, on which the standard radiomic features can be recomputed.

1.4 Feature selection

High-dimensional radiomics

As we saw in the previous section, modern radiomic pipelines extract a very large number of quantitative features from a single medical image. If we have multiple modalities, multiple VOIs, and/or if we use filters, then this number increases even further. Several hundred to several thousand features can be calculated for a single patient [34].

However, if the number of features greatly exceeds the number of observations:

$$N_{\text{features}} \gg N_{\text{patients}} \quad (1.2)$$

we face significant statistical challenges, collectively referred to as the *curse of dimensionality* [35, 36]. This concept, formalised in statistical learning theory, refers to the fact that in a very high-dimensional space, the volume of the space increases exponentially with the number of dimensions, which necessarily makes the data sparsely distributed [37]. In practical terms, entire regions of the feature space contain no observations. This sparsity prevents statistical learning models from generalising correctly, as they lack sufficient data points to reliably estimate the relationships between features and the target [37, 36]. To achieve equivalent statistical significance, high-dimensional spaces require a much larger number of patients than is typically available in radiomics studies [35].

Risks associated with high-dimensional data

There are several specific risks associated with this high-dimensional situation.

The first, and most significant, is overfitting. When a model has more parameters to estimate than there are observations, it can fit the training data perfectly, including its noise, without actually capturing the underlying patterns [36, 35]. Such a model then performs excellently on the training data but generalises poorly to new data. This risk is explicitly recognised in the radiomics literature as one of the main obstacles to clinical translation [9, 34].

The second risk is multicollinearity. Radiomic features are not independent: many features calculated from the same region and modality capture very similar information. The presence of highly correlated features can compromise the stability of model estimates, which may depend arbitrarily on one feature rather than another within a redundant group [30]. For this reason, reducing redundancy is an explicit goal of feature selection in radiomics [34].

Finally, there is a more practical constraint: computational cost, which increases rapidly with the number of features when training complex models [35].

Now that we have seen why we want to reduce the number of features, it is important to distinguish between two main categories of techniques for doing so:

1.4.1 Classification of selection methods

Feature selection methods are traditionally divided into three categories based on their relationship with the learning model used downstream [38, 39]:

Filter methods Filter methods rank features by calculating a score based on their intrinsic properties or their statistical relationship with the target, independently of any learning model [38, 39, 40]. Their main advantage is their low computational cost: no predictive model needs to be trained, making them particularly well-suited for high-dimensional data [39, 40]. They can then be used as an initial selection step before employing wrappers or embedded models [38]. Their main limitation is that they generally evaluate each feature independently of the others and therefore cannot detect interactions between features.

Wrapper methods Wrapper methods evaluate the relevance of a subset of features by training a predictive model on that subset and measuring its performance [38, 39, 40]. The process is iterative: we generate several candidate feature subsets, train a model on each one, and keep the one that yields the best predictive performance [39]. Their main advantage is their ability to account for interactions between features, which generally allows them to achieve better predictive performance. Their major drawback is their high computational cost, which often makes them inapplicable to very high-dimensional datasets [39, 40].

Embedded methods Embedded methods integrate feature selection directly into the training process of the predictive model [38, 40]. They represent a compromise between filter and wrapper methods, performing feature selection and model training in a single step [41]. They therefore have an intermediate computational cost. Their limitation is that they are specific to the model in which they are integrated [39].

1.4.2 Feature selection methods

The three feature selection methods used in this work were chosen to rely on distinct statistical principles, so that their combined use captures complementary aspects of feature relevance: a filter (univariate Cox), a model-agnostic importance measure (permutation importance), and an all-relevant algorithm (Boruta).

Univariate Cox

Univariate selection with the Cox model is a filter method commonly used in survival radiomics. A univariate Cox model (Section 1.5.3) is fitted to each feature taken in isolation, and the features are then ranked by a score reflecting the strength of their association with survival; the highest-ranked features are retained for subsequent analyses [42]. As a filter, the method is computationally cheap but evaluates each feature independently and therefore cannot capture interactions between features.

Permutation importance

Permutation importance is a method that measures the importance of a feature by quantifying the performance degradation of a predictive model when the values of that feature are randomly permuted [43]. The intuition is straightforward: if the model’s performance drops sharply after permutation, the feature is deemed important; if it does not change, the feature is considered non-informative. This method has the major advantage of being model-agnostic: it can be applied to any pre-trained predictive model. Its main limitation concerns the presence of correlated features: when two features carry similar information, the permutation of one is offset by the other, which artificially underestimates their respective importance [43].

Boruta

Boruta is an all-relevant feature selection algorithm. Unlike conventional approaches that seek the smallest feature set sufficient for prediction, Boruta aims to identify every potentially informative feature [44]. It relies on shadow features: for each original feature, a randomly permuted copy is created to serve as a pure-noise reference. A feature is deemed important if its importance score, computed by an ensemble model, typically a Random Forest, is significantly higher than that of the best shadow feature across repeated iterations. The algorithm classifies each feature as confirmed, tentative or rejected through an iterative statistical test. Its main limitation is its high computational cost, owing to the repeated training of the ensemble model at each iteration [44, 45].

1.5 Survival analysis

1.5.1 Principles of survival analysis

Survival analysis is used to model the time elapsed until a specific event occurs. The term “survival” is therefore used to represent this time until the event. This event can be the death of a patient, but also any discrete event, such as the time until a complication or the time until discharge from the hospital [46, 47]. The objective of survival analysis is to establish a relationship between the covariates and the time to the event. Survival data have two distinctive features, which make it impossible to use classical regression models [48].

Censoring In survival data, some observations are censored. Only certain individuals have experienced the event; the time to that event is unknown for the others. This can be due to various reasons, such as the study ending before the event occurred for some patients, or the loss of follow-up for patients. In these cases, the only available information is that the event had not yet occurred at the end of the study or at the time of the last update on the patient. If we do not take this information into account, we introduce a bias [49, 50].

For example, in Figure 1.8, we can see that:

- Patient A was lost to follow-up after three months without any events being recorded;
- Patients B and C experienced an event after enrollment, and before the end of the study;
- Patient D withdrew from the study two months after enrollment;
- Patient E did not experience any events before the end of the study.

Consequently, the exact timing of an event could only be recorded for Patients B and C; their records are not censored. For the other patients, it is unknown whether or not they experienced an event after the end of the study. The only valid information available regarding patients A, D, and E is that they had not experienced an event up to their last follow-up. Their data are therefore censored [50].

It should be noted that for standard methods of analysing survival data to be valid, censoring must be non-informative. That is, individuals who are censored must have the same probability of experiencing the event as those remaining in the study. Indeed, if individuals are removed from the study due to complications, the censoring has prognostic implications. Fortunately, if the proportion of censored patients is small, the bias is small and can be ignored [51, 52].

Skewness The second characteristic of survival data is that they do not follow a normal distribution. The time-to-event is, in most cases, relatively short, but for some patients, it can be very long. The statistical distribution of most durations exhibits a strong right-skewness. Models based on the assumption of a Gaussian distribution are therefore not applicable here [47].

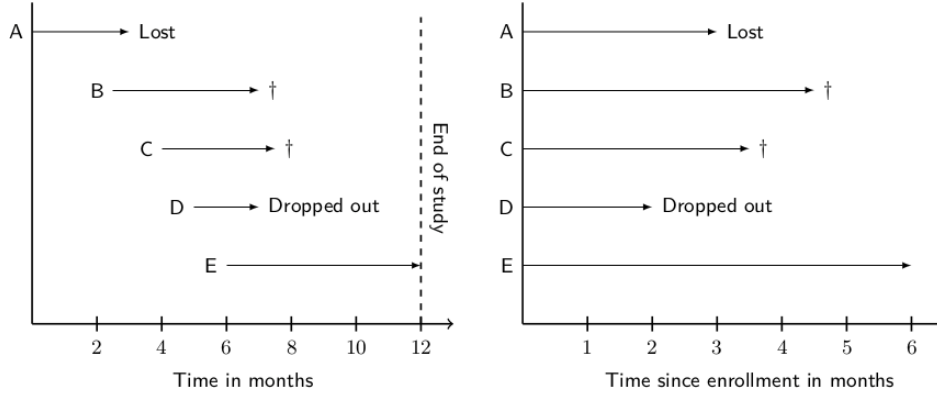


Figure 1.8: Illustration of right-censoring in a five-patient cohort, shown on the calendar timeline (left) and after realignment to the time since enrolment (right). Patient A was lost to follow-up, patients B and C experienced the event of interest ($\dagger$) during the study, patient D withdrew before the end of follow-up, and patient E remained event-free at the end of the study. Only B and C contribute exact event times; the other three observations are right-censored. Adapted from [50]

1.5.2 Functions used in survival analysis

As stated in Section 1.5.1, the two distinctive features of survival data, censoring and skewness, prevent us from predicting the event time T as a real-valued output of a classical regression. Survival analysis takes a different route: instead of predicting T itself, it characterises the distribution of T through dedicated functions. Two such functions, illustrated in Figure 1.9, lie at the heart of this work: the survival function $S(t)$, which describes the probability of being event-free beyond time t , and the hazard function $h(t)$, which describes the instantaneous risk of experiencing the event at time t , knowing that it did not occur until this time. Both characterise the same underlying distribution of T and are therefore mathematically equivalent, as explained later in this section, but they play complementary roles in the rest of this chapter: the Cox model and Gradient Boosting Survival target $h(t)$, the Random Survival Forest summarises $S(t)$ and the cumulative hazard $H(t)$ inside its leaves, and the evaluation metrics introduced in Section 1.5.4 ultimately rely on one or both.

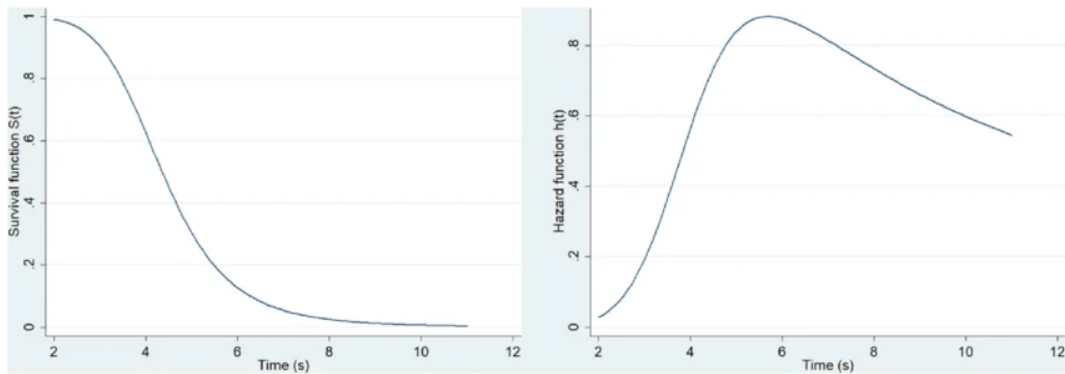


Figure 1.9: The two functions used to characterise a time-to-event distribution. The survival function $S(t)$ (left) is monotonically non-increasing, equals 1 at $t = 0$ and tends to 0 as t grows, and expresses the probability of being event-free beyond time t . The hazard function $h(t)$ (right) is a non-negative rate that describes the instantaneous risk of the event at time t , given that it has not yet occurred. Adapted from [53]

Before introducing them formally, we recall the notation used throughout the remainder of

this chapter [54]:

- T is a non-negative random variable representing an individual's survival time.
- t is a specific time of interest. The notation $T > t = 2$ therefore means that we are asking whether the individual has survived longer than 2 months, if the unit of time is the month.
- $d \in \{0, 1\}$ is the event indicator: $d = 1$ if the event was observed, in which case T is the actual event time, and $d = 0$ if the observation was censored, in which case T is only a lower bound on the unknown true event time.

Survival function $S(t)$ The survival function expresses the probability that an individual survives beyond a given time t :

$$S(t) = P(T > t). \quad (1.3)$$

It is a non-increasing function with $S(0) = 1$ (no one has experienced the event at the start of follow-up) and, under the assumption that the event eventually occurs for everyone, $S(t) \rightarrow 0$ as $t \rightarrow \infty$ [48, 50, 54]. $S(t)$ therefore summarises the survival experience of a study cohort.

As mentioned earlier, several of the evaluation metrics presented in Section 1.5.4 rely directly on $S(t)$, which makes it necessary to estimate this function from observed data. In the absence of censoring, the survival function can simply be estimated as the ratio of the number of patients surviving beyond time t to the total number of patients [50]:

$$\hat{S}(t) = \frac{\text{number of patients surviving beyond } t}{\text{total number of patients}}. \quad (1.4)$$

In the presence of censoring, however, this estimate becomes unreliable because the numerator is no longer known throughout follow-up: some patients exit the study before t without their final status being recorded. The **Kaplan-Meier (KM)** estimator avoids this issue by estimating survival incrementally at each observed event time, using only the patients still under observation at that moment [55, 48]:

$$\hat{S}_{\text{KM}}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}, \quad (1.5)$$

where $t_1, t_2, \dots$ are the distinct event times, d_i is the number of events at t_i , and n_i is the size of the at-risk set just before t_i , i.e. the number of individuals who are still being followed at that point [55, 48]. Censored patients contribute to n_i for as long as they are observed and then leave the at-risk set without counting as events, so their partial information is used while no assumption is made about their fate beyond censoring.

Graphically, $\hat{S}_{\text{KM}}(t)$ takes the form of a right-continuous step function that drops at each observed event time and remains flat between events. An example comparing two patient subgroups is shown in Figure 1.10. A typical use of KM curves is to stratify patients into groups, for instance, those receiving drug A versus drug B, and compare the resulting curves to assess whether the treatment has an impact on survival. The statistical significance of the difference between two such curves is then tested using a *log-rank test* [55, 56].

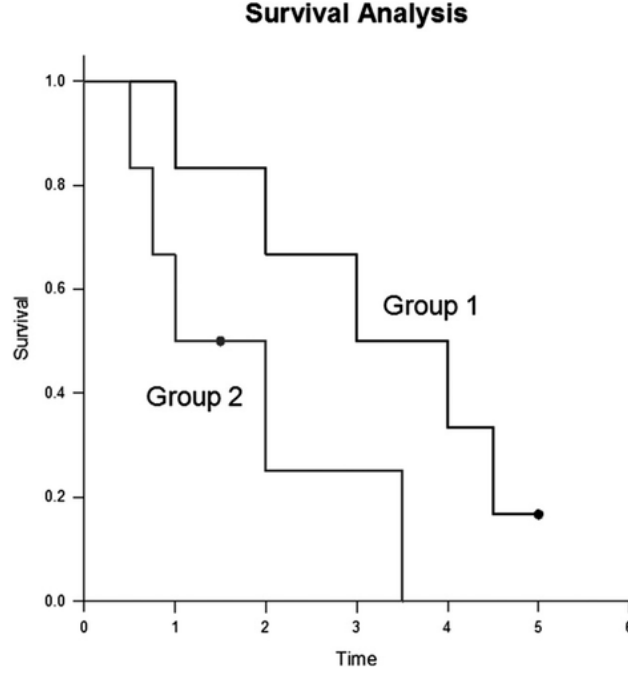


Figure 1.10: Kaplan-Meier estimates of the survival function for two patient subgroups followed over time. Each curve is a right-continuous step function that drops at every observed event and remains flat between events; the dots mark censored observations, which leave the at-risk set without contributing an event. Group 1 shows a higher survival probability than Group 2 throughout follow-up; the statistical significance of such a difference is typically assessed with a log-rank test. Adapted from [56]

Hazard function $h(t)$ The second function commonly used to characterise the distribution of T is the hazard function $h(t)$. Whereas $S(t)$ describes the cumulative likelihood of having survived up to time t , $h(t)$ describes the instantaneous risk of experiencing the event at time t , given that it has not yet occurred [54]:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}. \quad (1.6)$$

It is important to note that $h(t)$ is not a probability but a rate: it can take any non-negative value, and the product $h(t) \Delta t$ approximates the probability of experiencing the event in a short window $[t, t + \Delta t]$ for an individual still event-free at t . The interest of $h(t)$ lies in how it captures the time-varying nature of risk: it may increase with age, peak shortly after a treatment, or stay constant in a memoryless process. For this reason, $h(t)$ is the natural quantity to model when relating risk to patient covariates, as done by the Cox proportional hazards model and Gradient Boosting Survival (Section 1.5.3).

Cumulative hazard $H(t)$ Estimating $h(t)$ directly from data is difficult: since $h(t)$ is a rate, a nonparametric estimator would essentially amount to estimating a derivative from censored observations [54]. It is more convenient to work with its integral, the cumulative hazard function:

$$H(t) = \int_0^t h(u) du. \quad (1.7)$$

$H(t)$ is non-negative and monotonically increasing, and it can be estimated from censored data using the same incremental logic as the KM estimator. The resulting estimator is the

Nelson-Aalen (NA) estimator [54, 57]:

$$\hat{H}_{\text{NA}}(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}, \quad (1.8)$$

with the same notation as in Equation (1.5): at each observed event time t_i , the local event rate d_i/n_i is added to the running total. Like the KM estimator, the NA estimator will be used inside the leaves of the Random Survival Forest to summarise the cumulative risk of patients reaching a given leaf (Section 1.5.3).

Relationship between $S(t)$ and $h(t)$ The survival function $S(t)$ and the hazard function $h(t)$ are two equivalent representations of the same distribution of T : knowing one is enough to recover the other. Starting from the definition of $h(t)$ in Equation (1.6) and using the identity $f(t) = -dS(t)/dt$, where f is the probability density of T , the two functions can be linked through [54]:

$$h(t) = -\frac{d}{dt} \log S(t), \quad (1.9)$$

$$S(t) = \exp\left(-\int_0^t h(u) du\right) = \exp(-H(t)). \quad (1.10)$$

This equivalence has direct practical consequences for the rest of this work, since the survival models presented in Section 1.5.3 do not all output the same function. The Cox model and Gradient Boosting Survival target the hazard $h(t | X)$, whereas the Random Survival Forest produces nonparametric estimates of $S(t)$ and $H(t)$ in each leaf via the KM and NA estimators introduced above. Equations (1.9) and (1.10) guarantee that, whichever function a model natively outputs, the other can always be recovered afterwards. This conversion is essential at the evaluation stage (Section 1.5.4): the Brier score, for instance, compares predicted survival probabilities $\hat{S}(t | X)$ to observed outcomes, so models trained on the hazard side must convert their output to $\hat{S}(t | X)$ before being scored.

1.5.3 Survival Models

Section 1.5.2 introduced $S(t)$, $h(t)$ and $H(t)$ as descriptors of an unconditional time-to-event distribution. In practice, however, we want to relate this distribution to a vector of patient covariates $X = (X_1, \dots, X_p)$ so as to predict survival for new individuals. A survival model therefore estimates either the conditional hazard $h(t | X)$ or the conditional survival $S(t | X)$ from censored data. The duality established in Equations (1.9) and (1.10) guarantees that either function can be recovered from the other once one of them is known.

Five families of survival models are considered in this thesis. They differ in two main aspects: the function they natively model ($h(t | X)$, $S(t | X)$ or a relative risk score $f(X)$), and the structural assumptions they make on the relationship between X and risk.

Cox Proportional Hazards Model

The Cox Proportional Hazards (CoxPH) model expresses the conditional hazard as the product of a baseline hazard and a covariate-dependent exponential term [58, 59]:

$$h(t|X) = h_0(t) \cdot e^{\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p} \quad (1.11)$$

where $h_0(t)$ is the baseline hazard shared by all individuals, and $\exp(\beta^\top X)$ scales it by the patient's covariates.

The model is called semi-parametric because it makes no assumptions about the distribution of survival times, but does assume a specific form of the relationship between the covariates and the hazard function [58]. The functional form in Equation (1.11) carries a strong structural assumption known as the *proportional hazards assumption*: the hazard ratio between any two patients,

$$\frac{h(t | X_i)}{h(t | X_j)} = \exp(\beta^\top (X_i - X_j)), \quad (1.12)$$

is constant over time and depends only on their covariates [59]. Equivalently, the hazard curves of different individuals are scaled versions of one another and never cross. Under this assumption, the quantity $\text{HR} = \exp(\beta)$ admits a direct clinical interpretation as the **Hazard Ratio (HR)** associated with a one-unit increase in covariate X_k , all other covariates held fixed [58, 59]:

- $\text{HR} > 1$: X_k acts as a risk factor;
- $\text{HR} < 1$: X_k acts as a protective factor;
- $\text{HR} \approx 1$: X_k has no effect on the hazard.

The coefficients β are estimated by maximising the partial likelihood introduced by Cox [60, 61]:

$$L(\beta) = \prod_{i=1}^m \frac{e^{\beta^\top X_i}}{\sum_{j \in R_i} e^{\beta^\top X_j}}, \quad (1.13)$$

where R_i is the risk set at time t_i (the set of patients still under observation just before t_i). At each observed event, Equation (1.13) compares the patient who experiences the event to all those still at risk at the same time; the baseline hazard $h_0(t)$ cancels in this ratio, which allows it to remain unspecified. The same partial likelihood will serve as the building block for the penalised variants (Section 1.5.3) and as the loss function of Gradient Boosting Survival (Section 1.5.3).

Penalised Cox models

In the high-dimensional setting typical of radiomics, where the number of features p approaches or exceeds the number of patients n , the standard partial-likelihood maximisation becomes unstable: the optimum is poorly identified, coefficients explode, and the resulting model overfits (Section 1.4) [61]. Penalised variants restore stability by adding a regularisation term to the log partial likelihood. We consider the three classical variants [61, 62, 50]:

- **Cox-Ridge** (L_2 penalty):

$$\arg \max_{\beta} \log L(\beta) - \frac{\alpha}{2} \sum_{j=1}^p \beta_j^2. \quad (1.14)$$

The penalty shrinks all coefficients towards zero but never sets any of them exactly to zero. It controls multicollinearity while retaining the full set of features [50, 62].

- **Cox-LASSO** (L_1 penalty):

$$\arg \max_{\beta} \log L(\beta) - \alpha \sum_{j=1}^p |\beta_j|. \quad (1.15)$$

The geometry of the L_1 norm forces some coefficients to be exactly zero, so feature selection is performed jointly with estimation [50, 63].

- **Cox-ElasticNet**: combines the two penalties through a mixing parameter $r \in [0, 1]$

$$\arg \max_{\beta} \log L(\beta) - \alpha \left(r \sum_{j=1}^p |\beta_j| + \frac{1-r}{2} \sum_{j=1}^p \beta_j^2 \right). \quad (1.16)$$

Here $\alpha > 0$ controls the overall regularisation strength, while r balances the relative contributions of the two penalties: $r = 1$ recovers Cox-LASSO and $r = 0$ recovers Cox-Ridge. ElasticNet inherits the sparsity of LASSO and the better behaviour of Ridge in the presence of groups of correlated features [50, 62].

Because Cox-LASSO and Cox-ElasticNet drive part of β to zero during training, they implicitly select a subset of features [50]. They therefore belong to the family of embedded selection methods introduced in Section 1.4.1.

Random Survival Forest

Random Survival Forest (RSF) [64], extends the classical random forest to censored survival data. Unlike Cox models, the RSF is fully nonparametric: it makes no assumption about the form of $h(t | X)$, about the proportionality of hazards across patients, or about the linearity of the relationship between covariates and risk [50, 65].

The algorithm follows three steps [64, 65, 66, 50]:

1. **Bootstrap.** B bootstrap samples are drawn with replacement from the training set. On average, about 37 % of the original patients are left out of each sample and can be used for internal validation.
2. **Survival tree construction.** A survival tree is grown on each bootstrap sample. At each internal node, a random subset of covariates is drawn and, among them, the split chosen is the one that maximises the log-rank statistic between the two resulting child nodes, in order to favour splits that separate patients with markedly different survival.
3. **Aggregation.** Each terminal leaf summarises the survival experience of the patients it contains via the Kaplan-Meier estimator $\hat{S}(t)$ and the Nelson-Aalen estimator $\hat{H}(t)$ introduced in Section 1.5.2. For a new patient, the leaf reached in each of the B trees provides a leaf-level $\hat{S}$ and $\hat{H}$. The final predictions $\hat{S}(t | X)$ and $\hat{H}(t | X)$ are obtained by averaging these B leaf-level estimators.

The strength of the RSF is its ability to capture interactions and nonlinear effects automatically, without specifying any functional form [65].

Gradient Boosting Survival Models

Gradient Boosting [67] builds a predictive model as an additive sum of weak learners trained sequentially: each new learner is fitted to correct the residual errors of the previous ones [68]. This sequential construction distinguishes it from the Random (Survival) Forest, where trees are built independently and then averaged [50].

The survival extension keeps the same general logic but replaces the regression loss by the negative log partial likelihood of Equation (1.13), so that censored observations are handled correctly [69, 50]. At each iteration, a new weak learner is fitted to the negative gradient of this loss and added to the current additive function $F(X)$. The resulting hazard takes the form

$$h(t | X) = h_0(t) \cdot \exp(F(X)), \quad (1.17)$$

where $F(X) = \sum_m \nu g_m(X)$ is the cumulative contribution of the M weak learners g_m , weighted by the learning rate $\nu \in (0, 1]$ [50]. Equation (1.17) is identical in structure to the Cox model of Equation (1.11), the linear score $\beta^\top X$ being simply replaced by the flexible function $F(X)$: proportional hazards are still assumed (the time-dependence is still carried entirely by $h_0(t)$), but the dependency on X can be arbitrarily complex [69].

Two variants are considered in this work, distinguished by the choice of weak learner:

- **Tree-based Gradient Boosting Survival Models (GBSM).** Shallow regression trees are used as weak learners, which gives $F(X)$ a highly nonlinear and interaction-rich form [69, 50].
- **Gradient Boosting Survival Models Componentwise (GBSM CompWise).** The weak learners are univariate linear regressions: at each iteration, the model tests each feature individually and adds to F the one whose contribution decreases the loss the most [50]. The resulting $F(X)$ is therefore linear and sparse, in the same spirit as Cox-LASSO, and this variant is another embedded feature selection method (Section 1.4.1).

Survival Support Vector Machine

The Survival Support Vector Machine (SSVM) takes a different route from the previous models. Rather than estimating $h(t | X)$ or $S(t | X)$, it formulates survival prediction as a pairwise ranking problem [70, 50]: it learns a real-valued risk score $f(X)$ whose only purpose is to order patients correctly with respect to their event times. For two patients i and j such that $t_i < t_j$ and $d_i = 1$, the SSVM is trained so that $f(X_i) > f(X_j)$, violations being penalised by a squared hinge loss. Censoring is handled by restricting the loss to comparable pairs, that is, to pairs whose ordering can be determined from the observed data alone: a patient censored at time t_c therefore still contributes to all pairs in which another patient experienced the event before t_c , since we then know with certainty that the censored patient outlived that other one [70]. Only pairs whose ordering remains ambiguous, typically two censored patients, or a patient censored before another eventing patient, are absent from the loss.

In its linear form $f(X) = w^T X$, the goal is to minimise the objective function [71, 50]:

$$\arg \min_w \frac{1}{2} w^T w + \frac{\alpha}{2} \sum_{i,j \in \mathcal{P}} \max(0, 1 - (w^T x_i - w^T x_j))^2, \quad (1.18)$$

where $\mathcal{P} = \{(i, j) : t_i < t_j, d_i = 1\}$ is the set of comparable pairs and $\alpha > 0$ is a regularisation parameter.

Equation (1.18) assumes that the risk score is linear in the covariates. To allow nonlinear effects, the *kernel trick* [70, 50] replaces this linear score by a weighted sum of similarities between the new patient and the training observations:

$$f(X) = \sum_{k=1}^n \beta_k K(X_k, X), \quad (1.19)$$

where K , called a *kernel*, is any positive-definite similarity function on the covariate space.

Equation (1.18) makes the link to the concordance index explicit: the squared hinge loss is a convex surrogate for the count of misranked comparable pairs, and the C-index of Section 1.5.4 is exactly the proportion of correctly ranked ones. The SSVM is therefore the only model considered in this work that directly optimises a relaxation of its evaluation metric. Its main limitation is the converse of this strength: it returns only a relative score $f(X)$, with no associated $\hat{h}(t | X)$ or $\hat{S}(t | X)$, so the Brier score (Section 1.5.4), which requires absolute survival probabilities, cannot be computed for the SSVM.

1.5.4 Evaluation metrics

The five model families introduced in Section 1.5.3 produce predictions of different natures: a full $\hat{S}(t | X)$ for the Random Survival Forest, a full $\hat{h}(t | X)$ for the Cox model, its penalised variants and Gradient Boosting Survival, and a bare relative risk score $f(X)$ for the SSVM. A meaningful comparison therefore requires evaluation metrics that (i) accommodate censored data and (ii) are well defined across these output types.

Survival models are usually assessed along two complementary axes [50]: *discrimination*, the ability to rank patients correctly according to their risk, and *calibration*, the agreement between the absolute predicted probabilities and the observed event frequencies. The three metrics used in this work cover both axes: the concordance index gives a global, time-independent measure of discrimination; the Cumulative/Dynamic AUC refines it by resolving discrimination over time; and the Brier score adds the calibration dimension.

All three metrics must deal with the fact that, at any evaluation time t , some patients have been censored before t and their event status at t is therefore unknown. Naively dropping these patients introduces a bias, since censored patients are by construction those with shorter observed follow-up and are not representative of the cohort. Inverse-probability-of-censoring weighting (IPCW) [50] corrects this bias: each patient still informative at time t is reweighted by $1/\hat{G}(t)$, where $\hat{G}(t)$ is the Kaplan-Meier estimate of the probability of *not yet being censored* at t , computed by applying the estimator of Section 1.5.2 with the roles of event and censoring reversed [72]. A patient who had only a 50 % probability of remaining uncensored up to t therefore counts twice, in order to compensate, in expectation, for the patients with similar profiles who were censored and discarded. Both the Cumulative/Dynamic AUC and the Brier score below rely on this same weighting mechanism, the only difference between them is where $\hat{G}$ is evaluated.

Concordance Index (C-index)

The Concordance Index (C-index), also called Harrell's C-index, is a generalisation of the area under the Receiver Operating Characteristic (ROC) and measures the rank correlation between the predicted risk scores $\hat{f}(X)$ and the observed event times [50]. Reusing the comparable-pair definition used for SSVM (see Section 1.5.3), two patients i and j are *comparable* when $t_i < t_j$ and $d_i = 1$. Such a pair is *concordant* if the model assigns a higher risk score to the patient who experienced the event first, i.e. $\hat{f}_i > \hat{f}_j$. The C-index is the proportion of concordant pairs among the comparable ones:

$$\text{C-index} = \frac{\text{number of concordant pairs}}{\text{number of comparable pairs}} = \frac{\#\{(i, j) : t_i < t_j, d_i = 1, \hat{f}_i > \hat{f}_j\}}{\#\{(i, j) : t_i < t_j, d_i = 1\}} \quad (1.20)$$

It ranges from 0 to 1: C-index = 0.5 corresponds to random ranking and C-index = 1 to perfect discrimination [50].

Two well-known limitations are worth recalling [50]. First, the C-index is a global summary and does not reveal how discrimination varies with time. Second, Harrell's estimator in Equation (1.20) has been shown to become increasingly optimistic as the censoring rate grows: an IPCW alternative [73] corrects this bias by reweighting the comparable pairs to compensate for the censoring distribution.

Cumulative/Dynamic AUC

The Cumulative/Dynamic AUC ($\text{AUC}(t)$) is the time-resolved counterpart of the C-index. At each evaluation time t , it recasts survival as a binary classification problem [50]: patients who have experienced the event by t are the *cumulative cases* ($t_i \leq t$), and those still at risk are the *dynamic controls* ($t_i > t$). The area under the ROC curve for this two-class problem, denoted $\text{AUC}(t)$, then measures how well the model distinguishes patients who fail in $[0, t]$ from those who fail after t . As such, it is the appropriate metric when the clinical question is "what is the risk of an event *up to* time t ?" rather than "what is the risk *at* time t ?" [50].

Plotting $\text{AUC}(t)$ against t produces a time-resolved discrimination curve, which is strictly more informative than the global C-index: a model may discriminate well in the first months of follow-up and degrade afterwards as the at-risk set thins out, or vice versa [50]. The mean of $\text{AUC}(t)$ over a time window of interest provides a scalar summary close in spirit to the C-index, but which preserves the temporal dimension.

In practice, $\text{AUC}(t)$ is estimated using IPCW [50], and its variance grows quickly at the tails of the follow-up distribution, where the at-risk set or the number of events is small. The evaluation grid must therefore be restricted to a region with enough events and patients still at risk. Our choice for this work is detailed in Section 2.2.5.

Brier Score

Unlike the C-index and the time-dependent AUC, which only assess the ranking of predicted risks, the Brier Score (BS) evaluates the actual values of the predicted survival probabilities, making it a measure of both discrimination and calibration [50]. A well-calibrated model that

assigns a 90 % survival probability at time t to a group of patients should see approximately 90 % of them remain event-free at t .

At a given time t , the time-dependent Brier score is defined as [50]

$$\text{BS}(t) = \frac{1}{n} \sum_{i=1}^n \left[I(t_i \leq t \wedge d_i = 1) \frac{(0 - \hat{S}(t | \mathbf{x}_i))^2}{\hat{G}(t_i)} + I(t_i > t) \frac{(1 - \hat{S}(t | \mathbf{x}_i))^2}{\hat{G}(t)} \right], \quad (1.21)$$

where $\hat{S}(t | \mathbf{x}_i)$ is the survival probability predicted by the model for patient i at time t , and $\hat{G}(t)$ is the Kaplan-Meier estimate of the censoring distribution.

The score compares the predicted survival probability with the patient's true status at t , restricted to the two cases in which that status is unambiguous. In the first term, the patient experienced the event before or at t and is therefore known to be dead at t (true status 0); the squared deviation $(0 - \hat{S}(t | \mathbf{x}_i))^2$ penalises predictions that are too optimistic. In the second term, the patient is still under observation after t and is therefore known to be alive at t (true status 1); the squared deviation $(1 - \hat{S}(t | \mathbf{x}_i))^2$ penalises predictions that are too pessimistic. Patients censored before t have an unknown status at that time and contribute nothing to the sum.

This exclusion would bias the estimator, because patients with longer follow-up are also the most likely to have been censored. The inverse probability of censoring weights $1/\hat{G}(t_i)$ and $1/\hat{G}(t)$ correct for it: each retained patient is upweighted by the inverse of the probability that they would still be observed at the moment they contributed information, t_i for an event, t for a survivor, so that, in expectation, they stand in both for themselves and for the comparable patients lost to censoring.

In practice, integrating $\text{BS}(t)$ over the same time window used for the AUC yields the Integrated Brier Score (IBS), which summarises discrimination and calibration in a single scalar. Because Equation (1.21) requires an absolute survival probability $\hat{S}(t | \mathbf{x})$ at every evaluation time, the score cannot be computed for models that only produce a relative risk score; the SSVM (Section 1.5.3) is therefore excluded from this metric [50].

Restricted Mean Survival Time

Unlike the three metrics introduced above, which evaluate a model through its predicted risk scores or survival probabilities, the Restricted Mean Survival Time (RMST) summarises the survival experience of a cohort itself on an interpretable time scale. At a chosen horizon τ , the RMST is defined as the expected event-free time over the window $[0, \tau]$, which is mathematically equivalent to the area under the survival curve between 0 and τ [74, 75]:

$$\text{RMST}(\tau) = \mathbb{E}[\min(T, \tau)] = \int_0^\tau S(t) dt \quad (1.22)$$

In practice, $S(t)$ is replaced by its KM estimate (Equation (1.5)), and the integral is computed as the area under the resulting step function up to τ . The RMST is therefore expressed in the same units as the follow-up time (e.g. days or months) and admits a direct clinical reading as the mean event-free time experienced by a patient over the first τ units of follow-up.

Three properties make the RMST a useful complement to the metrics above [74]. First, its interpretation as a mean event-free duration is intuitive for both clinicians and patients,

whereas a hazard ratio or a C-index has no comparable physical meaning. Second, it reflects the whole survival history up to τ rather than survival at a single time point. Third, it remains a meaningful contrast between two groups even when the proportional hazards assumption fails, which is not the case for the hazard ratio derived from a Cox model. When two groups are compared, the difference of their RMSTs, denoted Δ -RMST, quantifies the average gain in event-free time of one group over the other within $[0, \tau]$ and is the natural effect-size measure on this scale.

1.6 State of the Art

The use of MRI-derived radiomics to predict OS in GBM has been an active research area for more than a decade. The present chapter follows a single thread: how the field has progressively moved from exploratory studies relying on very large, ad-hoc feature pools toward more compact, reproducible and clinically interpretable radiomic signatures. This trajectory directly motivates the methodological choices presented in Chapter 2.2.

Establishing radiomics as an independent prognostic biomarker

The first generation of GBM radiomic studies sought to demonstrate that quantitative imaging features could add prognostic information beyond what is already captured by age, the Karnofsky Performance Score (KPS)², and the extent of surgical resection. Kickingeder et al. (2016) [76] provided one of the earliest large-scale demonstrations of this principle on a cohort of 119 patients, extracting 12,190 features per patient from T1ce and FLAIR sequences and reducing them, via supervised PCA, to an 11-feature signature derived exclusively from the FLAIR modality. The radiomic model alone reached a C-index of 0.654 for overall survival, slightly outperforming a clinical model (0.640) and a radiological one (0.603) based on rCBV and Apparent Diffusion Coefficient (ADC), and the combination of radiomic and clinical variables reached 0.696. Although modest in absolute terms, these results were sufficient to position radiomics as a credible candidate biomarker.

A second wave of studies then asked whether this added value persisted once molecular variables were taken into account. Bae et al. (2018) [77], working on 217 patients with multiparametric MRI (T1ce, T2, FLAIR and DTI), extracted 796 features and used a RSF to integrate an 18-feature radiomic signature with clinical (age, KPS) and genetic (MGMT³, IDH1⁴) variables. The addition of the radiomic signature improved the integrated AUC for overall survival from 0.696 to 0.782, indicating that radiomic features capture prognostic information that is not redundant with established molecular markers. Kickingeder et al. (2018) [78] subsequently confirmed this on an independent 181-patient cohort using a Cox-LASSO model, finding that the radiomic signature and the MGMT status were the

²KPS is a clinical index ranging from 0 (death) to 100 (normal condition) that is used to assess a patient's functional capacity and level of independence in daily activities

³MGMT (O-6-methylguanine-DNA methyltransferase) is a DNA repair enzyme whose promoter methylation status serves as a key prognostic biomarker, as its inactivation improves the response of glioblastoma patients to temozolomide chemotherapy.

⁴HDH1 (isocitrate dehydrogenase 1) is a gene whose mutational status serves as a key molecular biomarker for the stratification of patients with glioblastoma, as its mutation is an independent predictor of a better prognosis compared with wild-type tumours.

only two variables to retain independent predictive value in a multivariate model, while the conventional tumour volume contributed nothing beyond the clinical baseline. By the end of the 2010s, the case for radiomics as an independent prognostic biomarker for GBM was therefore well established.

The BraTS era and the standardisation of public cohorts

A turning point came with the Multimodal Brain Tumor Segmentation Challenge (BraTS) challenges, which made multi-site, expert-segmented multimodal MRI cohorts publicly available and explicitly included an overall survival sub-challenge. Baid et al. (2020) [79], whose pipeline ranked third in the BraTS 2018 survival task, illustrate the methodology that became typical of this period: automatic 3D U-Net segmentation, extraction of 679 features from FLAIR and T1ce, reduction to 54 features through Spearman correlation and ANOVA, and classification of patients into short-, mid- and long-survival groups with a multilayer perceptron. Two findings stand out and reappear in most subsequent work: age is the single most important predictor, and the radiomic features extracted from the tumour core (necrosis and enhancing region) carry the bulk of the prognostic signal.

In parallel, several groups explored whether more physiologically informative sequences could push performance further. Park et al. (2020) [80] added diffusion-weighted (ADC) and perfusion-weighted (CBV) maps to conventional T1ce and FLAIR in 216 patients, extracted 6,472 features and selected six through Cox-LASSO regression, of which only three came from conventional sequences. Combined with age, KPS and extent of resection, this signature reached a C-index of 0.74 on the training cohort and 0.70 on an external validation cohort acquired with different protocols, an unusually robust result. While these findings confirm the prognostic value of advanced sequences, diffusion and perfusion are not routinely acquired in every centre and are absent from the BraTS cohorts; for studies aiming at broad clinical applicability, the four conventional T1, T1ce, T2 and FLAIR sequences therefore remain the natural substrate.

Deep radiomics: gains in performance, losses in interpretability

A parallel line of work has examined whether features generated by deep neural networks can outperform the hand-crafted features computed by standardised pipelines such as PyRadiomics. Lao et al. (2017) [81] extracted 98,304 deep features by transfer learning from a CNN pre-trained on natural images and selected six of them through a Cox-LASSO model on 112 GBM patients, reaching a C-index of 0.710 on a validation cohort, which rose to 0.739 once age and KPS were added. Chaddad et al. (2020) [82] reported a similar pattern on 100 recurrent GBM patients, with deep radiomic features reaching an AUC of 0.89 against 0.78 for matched hand-crafted features. More recently, Zhang et al. (2023) [83] compared CoxPH, SurvivalTree, RSF, DeepSurv and DeepHit on 131 patients and found DeepSurv to be the best performer at a C-index of 0.732, ahead of the RSF at 0.667 and the classical CoxPH at 0.635. The picture that emerges is that deep approaches can extract additional signal, but at the cost of two practical drawbacks: a much larger sample size requirement, which remains a clear limitation in GBM cohorts of a few hundred patients, and a loss of biological interpretability of the resulting features, which complicates clinical translation.

Toward parsimony, reproducibility and interpretability

The most recent studies have largely turned away from the race to ever-larger feature pools and instead emphasise the construction of compact, reproducible and interpretable signatures. Hajianfar et al. (2023) [84] systematically explored 504 combinations of preprocessing, feature-selection and machine-learning algorithms on 119 patients and reported a best C-index of 0.77, achieved with a Laplacian-of-Gaussian filter, mutual-information selection and a Cox Boost estimator. Beyond this headline figure, the value of their work lies in the explicit demonstration that the choice of selection method and modelling family materially affects performance, and that no single combination dominates across configurations. Styliara et al. (2024) [42] followed a similar philosophy but with a focus on biological interpretation: working on 49 IDH-wildtype GBM patients, they segmented four tumour sub-regions and used PyRadiomics to extract 1,862 features, eventually reducing the model to age and six radiomic features via Cox-LASSO, and providing a tissue-level interpretation for each selected descriptor.

The clearest illustration of this shift is the recent work of Duman et al. (2024) [85]. Their study draws on 289 patients, including the entire BraTS2020 cohort and explicitly aims at IBSI-compliant, reproducible features and a minimal model size of three to ten variables. After a stability analysis, the final signature contains only three variables: age and two FLAIR-derived radiomic features (a surface-to-volume morphology descriptor and a grey-level zone non-uniformity texture descriptor). Despite this extreme parsimony, the model reaches a C-index of 0.69 on the validation cohort and an integrated AUC at 11 months of 0.81, reported as the highest in the GBM survival-radiomics literature at the time of writing. Duman et al. thus demonstrate that a small, transparent feature set extracted from a single conventional modality can match the performance of much larger and more complex pipelines, provided that stability and reproducibility are enforced from the outset.

Positioning of the present work

Three observations can be drawn from this review and structure the methodology developed in Chapter 2. First, recent GBM studies based on cohorts of a few hundred patients have predominantly relied on hand-crafted features extracted with standardised pipelines such as PyRadiomics rather than on deep-learning features: the performance gains reported for deep features are accompanied by larger sample-size requirements and by reduced biological interpretability [83], two aspects repeatedly noted as limitations in cohorts of this size. The present work follows this line and is restricted to hand-crafted features extracted from the four conventional sequences (T1, T1ce, T2 and FLAIR) available in most clinical protocols. Second, performance has been shown to depend on the choice of survival algorithm: Hajianfar et al. (2023) [84] and Zhang et al. (2023) [83] reported that different algorithmic families extract different signals from the same feature set, with no single configuration dominating across settings. Rather than selecting a single estimator in advance, this thesis therefore evaluates several survival-model families under a common protocol. Third, age has repeatedly been reported as the single most important predictor of overall survival in this literature. As the BraTS2020 cohort provides no molecular variables and only age as a usable clinical covariate, age is the sole clinical input of the model proposed by Duman et al. (2024) [85] on this cohort and is likewise retained here alongside the radiomic features. Conducted on the same BraTS2020 cohort as Duman et al., the present work is therefore directly comparable

to a recent benchmark in the GBM survival-radiomics literature, from which it differs by evaluating several survival-model families rather than relying on a single one.

Chapter 2

Materials and Methods

2.1 Materials

The experiments carried out in this work relied on the dataset from the Multimodal Brain Tumor Segmentation Challenge 2020 (BraTS2020), jointly organised by the center for Biomedical Image Computing and Analytics (CBICA) at the University of Pennsylvania and the Medical Image Computing and Computer Assisted Intervention (MICCAI) conference [11, 12, 13, 14, 15].

Cohort composition The BraTS2020 dataset gathers pre-operative MRI examinations acquired during routine clinical care from patients diagnosed with diffuse gliomas, namely glioblastomas and lower-grade gliomas, with histologically confirmed diagnosis. The data were collected across 19 different institutions, which introduces substantial variability in clinical protocols and scanners. Since this work focuses on survival analysis of GBM patients, we describe only the portion of the data that is relevant to this work. Out of all the cases available in this dataset, 236 patients had GBM and an associated Overall Survival (OS) value in days. These patients all belong to the training subset. We nevertheless worked with 234 patients, as those for whom at least one of the three tumour sub-regions was absent from the segmentation were excluded, as radiomic features cannot be extracted from an empty sub-region. The characteristics of this cohort are summarised in Table 2.1.

Characteristic	Value
Patients, n	234
Age (years), median [range]	61.5 [19.0–86.7]
OS (days), median [range]	369 [5–1767]
Events observed, n	233
Censoring rate, %	0.4

Table 2.1: Baseline characteristics of the BraTS2020 GBM cohort used in this work ($n = 234$ patients with available OS information).

Imaging modalities For each patient, four 3D MRI sequences are provided, which form the input modalities of the segmentation problem [11, 12, 13]:

- T1: T1-weighted;
- T1ce: T1-weighted after injection of a gadolinium-based contrast agent;
- T2: T2-weighted;

- FLAIR: T2 Fluid-Attenuated Inversion Recovery.

Each volume is distributed in NIfTI format (.nii.gz) with a uniform size of $240 \times 240 \times 155$ voxels.

Preprocessing applied by the organisers The images are provided after a standardised preprocessing pipeline applied by the organisers, which guarantees consistency between subjects and across institutions [25]:

- Conversion of DICOM files to NIfTI format;
- Co-registration of the four modalities of a given patient to a common anatomical reference atlas (SRI24);
- Isotropic resampling to a resolution of 1 mm^3 per voxel;
- Skull-stripping (brain extraction), which removes non-cerebral structures (skull, skin, fat) that could otherwise introduce bias during learning.

This preprocessing had an important consequence for the rest of this work: no inter-modality registration, resampling, or brain extraction step needed to be performed in our own pipeline, since these operations were already carried out upstream. Intensities, however, were not normalised between patients, which made a dedicated normalisation step necessary in our processing chain, as discussed in Section 2.2.1.

Annotations and tumour sub-regions The reference annotations (ground truth) were produced manually by one to four operators following a common annotation protocol, and were then validated by experienced neuro-radiologists [11, 12, 13]. Three elementary tumour classes are annotated voxel by voxel:

- NCR/NET (label 1): necrosis and active non-enhancing tumour;
- ED (label 2): peritumoural oedema;
- ET (label 4): tumour enhancing after gadolinium injection.

In practice, the BraTS2020 dataset provides annotations using three labels rather than four: necrosis and non-enhancing tumour are merged into a single combined label. This merger, introduced in BraTS from 2017 onwards, reflects the difficulty of reliably distinguishing pure necrosis from non-enhancing tumour tissue on standard MRI modalities [25].

Associated clinical data In addition to the images, a CSV file accompanies the dataset and provides, for each pseudonymised patient, the age, the OS in days, and the resection status. The latter was not used in this thesis. Among the 234 patients retained, 118 are labelled Gross Total Resection (GTR), 9 are labelled STR (Subtotal Resection), and 107 have no resection status reported (NA). Restricting the analysis to GTR patients, as the BraTS2020 survival sub-challenge does, would shrink the cohort by half and discard most of the available survival information. Since the resection status is missing for the majority of patients, we used age as the only clinical covariate and let the radiomic features carry the rest of the prognostic signal.

2.2 Methods

The methodological pipeline of this work followed the five sequential stages of the radiomics workflow introduced in Section 1.3: image preprocessing, feature extraction, feature selection, model selection, and final evaluation. Each stage is described in a dedicated subsection below. All experiments were carried out in `Python` (3.10.8), with `PyRadiomics` (3.0.1) for feature extraction, `scikit-survival` (0.25.0) for survival modelling, `scikit-learn` (1.7.2) for cross-validation (CV) and standardisation and `BorutaPy` for the implementation of the Boruta selection method.

2.2.1 Image preprocessing

As stated in Section 2.1, the data we used had already undergone several preprocessing steps applied in a uniform manner by the BraTS challenge organisers. This pipeline harmonised the images geometrically, but it did not correct inter-patient intensity inhomogeneity. Unlike CT, MRI is fundamentally a qualitative modality: the intensity values of a voxel have no absolute physical meaning and depend on the scanner manufacturer, the magnetic field strength, the sequence parameters, and the spatially varying sensitivity of the receiver coils [86]. For radiomics, this variability constitutes a major obstacle to the stability of the extracted features and to the generalisation of models across cohorts [86, 87].

Figure 2.1 illustrates this phenomenon concretely on our dataset: before normalisation (left panel), the T1 intensity distributions of six representative patients differed considerably, with modes ranging from approximately 300–500 for some patients to 1000–1500 for others, although they all originated from anatomically comparable brain examinations. Since the BraTS2020 cohort was collected across 19 different institutions, this effect is particularly likely to affect downstream analyses. We therefore added an intensity standardisation step based on the z-score, a classical method widely used in brain MRI radiomics that offers a trade-off between simplicity, reproducibility, and effective harmonisation of inter-patient distributions [86].

Normalisation was applied to the 234 patients retained according to the inclusion criteria described in Section 2.1, and independently to each of the four modalities (T1, T1ce, T2, FLAIR). For each volume, a brain mask was defined as the set of voxels with non-zero intensity, which constituted a reliable proxy for the true brain mask, given that the skull-stripping applied upstream by BraTS set all extra-cerebral voxels to zero. The mean μ and the standard deviation σ were then computed exclusively on the voxels belonging to this mask, so that background voxels would not bias the statistics by artificially lowering the mean and inflating the standard deviation. Each intra-mask voxel of initial intensity x was then transformed according to:

$$x' = \frac{x - \mu}{\sigma + \epsilon} \quad (2.1)$$

where $\epsilon = 10^{-8}$ is a numerical regularisation term that prevented division by zero. Extra-cerebral voxels were preserved at 0 in order to maintain the consistency of the brain mask, and the tumour segmentation masks were not modified. The normalised images were saved in NIfTI format while preserving the original affine and header, which guaranteed geometric consistency with the segmentation masks used later for feature extraction.

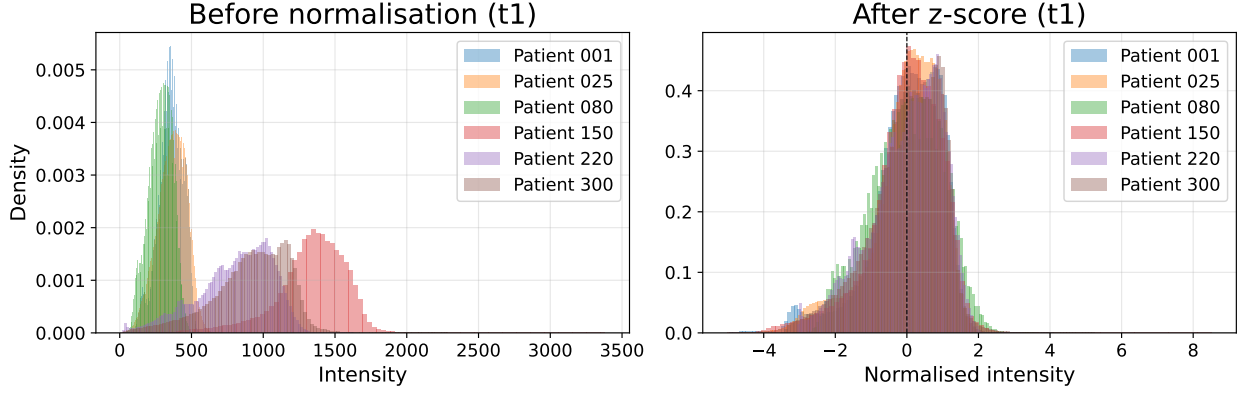


Figure 2.1: Distribution of T1 intensities for six representative patients of the cohort, before (left) and after (right) z-score normalisation. Before normalisation, the distributions exhibit strong inter-patient heterogeneity, with modes ranging from approximately 300–500 (patients 001, 025, 080) to 1000–1500 (patients 150, 220, 300), reflecting the variability inherent to multi-centre MRI acquisitions. After z-score normalisation, applied independently to each volume, the distributions are centred and scaled and overlap almost perfectly, demonstrating the effective harmonisation of intensities across patients.

It is worth emphasising that the statistics μ and σ were computed intra-image: each volume was normalised using its own statistics only, without resorting to global statistics aggregated over the entire dataset. This choice had two advantages. First, it prevented any data leakage between the future training, validation, and test subsets (see Section 2.2.3), since no statistic from the full dataset entered into the transformation of any given image. Second, it made the procedure directly transferable to a new patient at inference time, without the need to recompute cohort statistics. The trade-off was that the absolute meaning of intensities was lost after normalisation, which is consistent with the qualitative nature of MRI and does not affect the texture features.

The effect of this normalisation is illustrated quantitatively and qualitatively in Figures 2.1 and 2.2. In Figure 2.1 (right panel), the intensity distributions of the six patients, initially highly dispersed, overlap almost perfectly after normalisation, centred on 0 with a standard deviation close to 1. Figure 2.2 shows an axial T1 slice of patient 001 before and after normalisation: although the absolute intensity scale changes (from [0, 550] to approximately [4, 2]), the local anatomical contrast and the structure of the brain tissues are entirely preserved, confirming that the transformation is purely a rescaling and does not alter the radiological information.

No bias field correction (e.g. N4ITK [88]) was applied prior to z-score standardisation. The choice and its trade-offs are discussed in Chapter 4.5.

2.2.2 Feature extraction

Radiomic feature extraction was carried out using the open-source Python library `PyRadiomics` [32], which provides an implementation of radiomic feature definitions largely aligned with the IBSI consensus, with a few documented deviations [89]. We deliberately restricted ourselves to hand-crafted features, excluding deep-learning-derived features, in order to preserve the biological interpretability of every variable retained downstream (see Section 1.3).

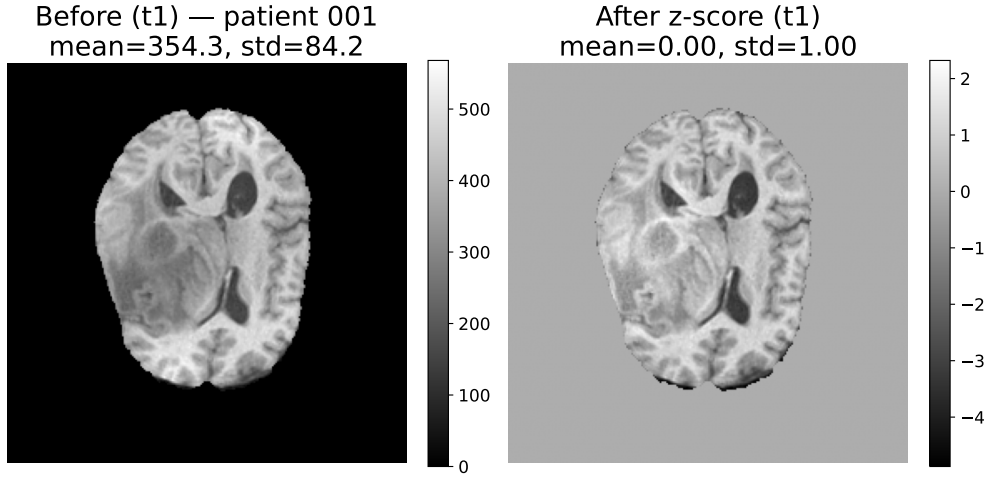


Figure 2.2: Axial T1 slice of patient 001, before (left) and after (right) z-score normalisation. Before normalisation, intensities range from 0 to approximately 550 (mean = 354.3; standard deviation = 84.2). After normalisation, the intra-cerebral intensities are centred on 0 with a standard deviation of 1, while fully preserving the local anatomical contrast and the differentiation between cerebral tissues. Extra-cerebral voxels remain at 0 (displayed as uniform grey on the right panel due to the rescaling of the colour bar).

All seven feature classes available in `PyRadiomics` were enabled, in order to maximise the richness of the tumour characterisation. The theoretical content of each class is described in Section 1.3.3. We recall here only the number of features extracted per class in `PyRadiomics`: First Order Statistics (19), Shape-based 3D (17), GLCM (24), GLRLM (16), GLSZM (16), GLDM (14), and NGTDM (5). The complete list of features extracted is provided in Table A.1.

For every modality, features were extracted both on the original image and on images derived from a 3D discrete wavelet decomposition (`enableImageTypeByName("Wavelet")`). The decomposition was performed with the `PyRadiomics` defaults: a first-order Coiflet mother wavelet (`coif1`) and a single decomposition level, yielding the eight sub-bands whose construction is described in Section 1.3.4. This gave nine image representations per modality (one original plus eight wavelet sub-bands), an enrichment commonly used in brain-tumour radiomics to capture textural heterogeneity at several spatial scales and orientations [90, 91].

To exploit intratumoural heterogeneity, features were extracted independently from each of the three tumour sub-regions annotated in BraTS2020 (see Section 2.1): the necrotic and non-enhancing tumour core (label 1, NCR), the peritumoural oedema (label 2, ED), and the GD-enhancing tumour (label 4, ET). Intensity- and texture-dependent features are sensitive to the underlying voxel values and were therefore computed separately on each of the four modalities (T1, T1ce, T2, FLAIR). Shape features, which depend only on the segmentation mask, are by construction intensity-independent and were computed only once per sub-region.

In addition to the radiomic features, a single non-imaging variable was included in the feature vector: the patient’s age at diagnosis, available in the BraTS2020 dataset and a well-established prognostic factor for high-grade gliomas [84].

The full feature vector per patient was obtained as the product of the independent dimensions described above. The 94 intensity- and texture-based features ($19 + 24 + 16 + 16 + 14 + 5$) were computed for each of the 4 modalities, each of the 9 image types and each of the 3

sub-regions. The 17 shape features were computed only once per sub-region. Adding the patient’s age finally yielded:

$$N_{\text{features}} = \underbrace{4}_{\text{MRI modalities}} \times \underbrace{9}_{\text{image types}} \times \underbrace{94}_{\text{intensity \& texture}} \times \underbrace{3}_{\text{sub-regions}} + \underbrace{17 \times 3}_{\text{shape per sub-region}} + \underbrace{1}_{\text{age}} = 10\,204. \quad (2.2)$$

A summary of these feature counts is given in Table 2.2. To make every feature traceable throughout the downstream pipeline, we adopted the naming convention `modality_filter_class_featureName_zone`, with the single exception of the clinical variable `age`. For example, `t1ce_wavelet-LLH_glcml_Contrast_NCR` denotes the GLCM Contrast feature extracted from the LLH wavelet sub-band of the T1ce image, in the necrotic core sub-region.

Table 2.2: Summary of the calculation of the total number of features extracted per patient.

Class	n_{class}	Modalities	Image types	Zones	Total
First Order	19	4	9	3	2 052
GLCM	24	4	9	3	2 592
GLRLM	16	4	9	3	1 728
GLSZM	16	4	9	3	1 728
GLDM	14	4	9	3	1 512
NGTDM	5	4	9	3	540
Shape 3D	17	1	1	3	51
Subtotal : radiomic features					10 203
Clinical (age)	1	/	/	/	1
Total per patient					10 204

2.2.3 Feature selection

The extraction described in Section 2.2.2 yielded a vector of 10,204 predictors per patient for 234 patients. With a predictor-to-observation ratio of approximately 44, direct estimation of a multivariate model is impossible without a prior dimensionality-reduction step. This master’s thesis relied exclusively on feature selection methods, and no dimensionality reduction methods, in order to preserve the biological interpretability of the selected radiomic features. This approach allows the selected features to be directly linked to the tumour characteristics they measure, which facilitates the clinical interpretation of the final model. Feature selection here pursued two objectives: to eliminate redundant or non-informative features and to limit the risk of overfitting downstream. The procedure was organised in four stages: (i) train/test split, (ii) unsupervised preliminary filter, (iii) multi-method supervised selection, (iv) inter-method consensus.

(i) Train/test split

Before any selection step, the patient set was partitioned into a training set (187 patients, 80%) and an independent test set (47 patients, 20%). All the steps described below were

performed exclusively on the training set, the test set remaining strictly isolated until final evaluation (Section 2.2.5).

As shown in Table 2.3, the censoring rate in this cohort was extremely low: a single patient is censored in the training set (0.5%) and none in the test set. This particularity had two consequences. On the one hand, survival analysis retained its formal relevance, survival models remained applicable and their metrics interpretable, but the specific richness of censored data, which historically justifies recourse to these models rather than to a direct regression on event time, was here barely mobilised. On the other hand, the log-rank test between train and test ($p = 0.598$) revealed no significant difference in survival distributions, which guaranteed the representativeness of the split but also reflected the low variability induced by the near-total absence of censoring.

Table 2.3: Description of the BraTS2020 GBM cohort. Continuous variables are reported as median [Interquartile Range (IQR)] and [range]. P-values were computed using the Mann–Whitney U test for continuous variables and the log-rank test for survival distributions.

Variable	Total	Train	Test	p-value
n	234	187	47	—
Age, median [IQR]	61.5 [54.3–69.2]	62.2 [54.4–69.4]	59.0 [54.1–67.2]	0.222
Age, range	[19.0–86.7]	[19.0–86.7]	[36.8–80.5]	
OS days, median [IQR]	369 [192–578]	376 [196–578]	334 [190–558]	0.301
OS days, range	[5–1767]	[12–1767]	[5–1592]	
Events observed	233	186	47	0.598
Censoring rate (%)	0.4	0.5	0.0	
Log-rank train vs. test	—	—	—	0.598

(ii) Unsupervised preliminary filter

A single unsupervised filter was applied to reduce dimensionality before supervised selection. This filter is termed unsupervised because it uses neither the target variable nor the censoring indicator: only the feature values across patients are considered. Its role was to reduce the most pronounced redundancy among radiomic features before engaging the more costly supervised methods.

A redundancy filter based on the absolute Spearman correlation was applied: for any pair of features (X_i, X_j) such that $|\rho(X_i, X_j)| > 0.75$, only one of the two was retained.

This filter was motivated by the strong intrinsic redundancy of radiomic features: many features extracted from the same region and modality capture very similar information. The threshold of 0.75 is consistent with practices reported in the radiomics literature, where thresholds between 0.7 and 0.8 are commonly used [42, 92]. A stricter threshold (e.g. 0.9) would have left considerable redundancy untouched and unnecessarily inflated the supervised selection step. A more permissive threshold (e.g. 0.6) would have risked prematurely discarding features carrying complementary information. The chosen threshold therefore represented a standard compromise.

When two features exceeded the correlation threshold, the one retained was determined by alphabetical order of feature names: the feature whose name came first alphabetically was kept, and the other discarded. This rule is purely conventional and carries no statistical justification, it is documented here as a methodological limitation.

(iii) Multi-method supervised selection

Three supervised selection methods were applied in parallel to the features retained after the redundancy filter: a univariate Cox filter, permutation importance, and the Boruta algorithm. Because each method rests on a distinct statistical principle, their joint use captures complementary aspects of feature relevance and allows the robustness of the selection to be assessed with respect to the chosen method. The principles of the three methods are presented in Section 1.4.2.

Univariate Cox screening A univariate Cox model (Section 1.5.3) was fitted to each feature taken in isolation on the training set [50], and its discriminative ability was measured by the C-index computed on the same data. Features were then ranked by decreasing $|\text{C-index} - 0.5|$; this transformation retained features associated with the risk either positively or negatively, the sign being resolved by the models downstream.

Permutation importance A RSF was first fitted on the full set of features. For each feature, its values were randomly permuted across patients, the C-index was recomputed on the (non-retrained) model, and the resulting drop in performance was taken as the importance score. The procedure was repeated thirty times per feature to mitigate the variance arising from permutation randomness, and the mean scores were used to rank the features. The RSF was chosen as the base estimator for three reasons: (1) it natively handles censored data through the log-rank statistic in its node-splitting criterion (Section 1.5.3), which made it directly compatible with our data without modification; (2) it makes no proportional-hazards or linearity assumption, allowing non-linear relationships between radiomic features and survival to be captured; and (3) it is consistent with the dimensionality considered here. Its parameters are summarised in Table 2.4. The known attenuation of permutation importance in the presence of correlated features was partly mitigated here by the upstream Spearman filter, though not eliminated [43].

Table 2.4: Parameters of the RSF used for permutation importance.

Parameter	Value
n_estimators	100
max_features	$\sqrt{p}$
min_samples_leaf	10
random_state	42
Number of permutations	30
Metric evaluated	C-index

Boruta Boruta was run with a RSF as its internal estimator. Although the algorithm was originally designed for classification or regression with a Random Forest, its application to survival data requires an estimator that produces a per-feature importance score while respecting censoring; the RSF is the natural survival-analysis counterpart of the Random Forest and the standard choice for this adaptation [44]. Using a RSF (i) preserves censoring information in tree construction and (ii) aligns the selection with the same family of inducers as the evaluation metric used downstream. The algorithm was applied through the **BorutaPy** library [93], with the parameters summarised in Table 2.5. **BorutaPy** classified each feature as confirmed, rejected, or tentative; to remain inclusive at this stage, and given the downstream inter-method consensus step, both confirmed and tentative features were retained.

Table 2.5: Parameters of the Boruta algorithm and of the internal RSF used as the estimator.

Parameter	Value
<i>BorutaPy</i>	
α	0.05
two_step	True
perc	100
max_iter	100
n_iter_no_change	20
<i>Internal RSF</i>	
n_estimators	100
max_features	$\sqrt{p}$
min_samples_leaf	10
random_state	42

The two methods that produce a ranking (univariate Cox and permutation) require an explicit choice of the number k of features to retain. Boruta, by contrast, directly produces a subset and requires no optimisation of k : the selected features were passed as-is to the downstream models.

For each method $\times$ model combination (except Boruta), we swept the number of features k from 1 to 50 in steps of 1, adding at each step the first k features of the ranking to the model. Each value of k was evaluated by 5-fold Cross-Validation (CV) within the training set: at each fold the model was fitted on four folds and scored on the remaining held-out fold, and the cross-validated C-index for a given k was the mean of the five held-out scores, with its standard deviation taken across the folds. The upper bound $k_{\max} = 50$ was retained as a compromise between exhaustiveness of the sweep and computational cost: beyond 50 features for 187 patients, the risk of overfitting outweighs the expected marginal predictive gain.

Rather than retaining the value of k that strictly maximised the cross-validated C-index, we selected the number of features with a rule inspired by the one-standard-error (1-SE) rule of Hastie et al. [94]. Among all values of k whose mean cross-validated C-index laid within a tolerance of the maximum, we retained the smallest one. We defined this tolerance as One Standard Deviation (1-SD) of the cross-validated C-index across folds (at the maximising k), rather than the standard error of the mean used in the original formulation. Since the standard deviation exceeds the standard error by a factor $\sqrt{K}$ (here $\sqrt{5}$), this widens the equivalence band and therefore yielded a more conservative, more parsimonious selection. This choice was motivated by the high dimensionality of the radiomic feature space, where parsimony,

interpretability, and a reduced risk of overfitting are particularly valuable; performances lying on the resulting plateau of the C-index-versus- k curve were in any case not statistically distinguishable from the maximum. Figure 2.3 illustrates this procedure for the GBSM CompWise $\times$ permutation-importance combination.

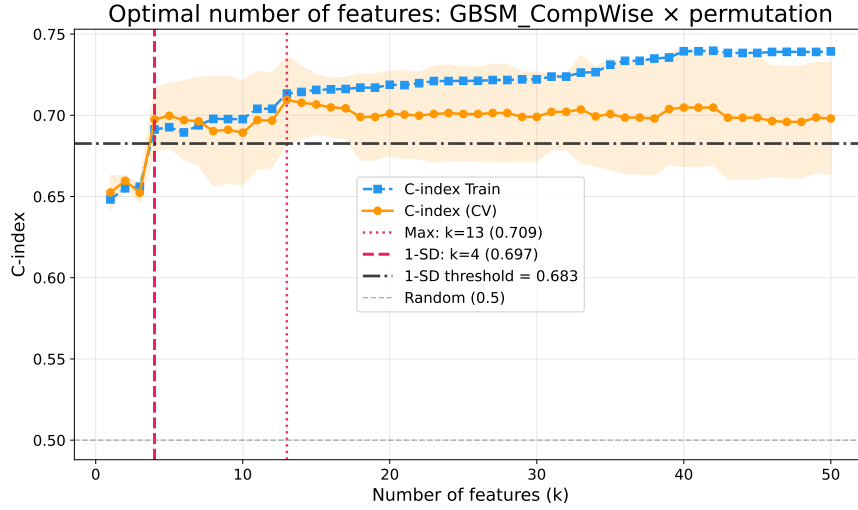


Figure 2.3: Illustration of the one-standard-deviation rule for the GBSM CompWise $\times$ permutation-importance combination. The orange curve is the 5-fold cross-validated C-index (mean across held-out folds, computed within the training set) as a function of the number of retained features k , with the shaded band showing ± 1 standard deviation across folds; the blue curve is the corresponding training C-index, whose gap to the orange curve reflects overfitting. The dotted vertical line marks $k^* = 13$, the value maximising the mean cross-validated C-index (0.709); the dashed vertical line marks the retained $k_{1\text{-SD}} = 4$, the smallest k whose mean cross-validated C-index exceeds the threshold $C(k^*) - \text{SD}(k^*) = 0.683$. The horizontal dashed line at 0.5 marks the random-ranking baseline.

Each selected feature set was combined with eight survival models covering a broad range of statistical assumptions. The choice of these models is detailed in Section 2.2.4:

- Cox proportional hazards (CoxPH)
- Cox-Ridge
- Cox-LASSO
- Cox-ElasticNet
- Random Survival Forest (RSF)
- Survival Gradient Boosting (GBSM)
- Component-wise Survival Gradient Boosting (GBSM CompWise)
- Survival Support Vector Machine (SSVM)

All models were implemented via the `scikit-survival` library [95]. Models with internal selection (Cox-LASSO, Cox-ElasticNet) remained free to retain only a subset of the features passed to them, through their L_1 penalisation 1.5.3.

In total, 24 combinations were thus evaluated (8 models $\times$ 3 selection methods).

(iv) Final selection by consensus

For each feature selected at least once by one of the three supervised methods, we counted the number of (method $\times$ model) combinations that retained it. This count provided a robustness indicator: features selected consistently by several independent approaches were more likely to capture genuine signal than those whose selection depends on a single combination.

The final features retained for modelling were those selected in at least two of the 24 (method $\times$ model) combinations. Each combination constituted a distinct analytical configuration, pairing one of the three selection methods (univariate Cox filter, permutation importance, Boruta) with one of the eight survival models considered. The ≥ 2 threshold was motivated by the following considerations:

- Each configuration carried its own biases. The three methods rely on distinct statistical principles and each has known limitations: a univariate filter ignores interactions, permutation importance under-estimates correlated features, and Boruta is sensitive to the parameterisation of its statistical test. The eight survival models, in turn, impose different structural assumptions on the data. A feature retained in at least two configurations was therefore less likely to be an artefact of one specific method-model interaction than a feature selected in a single configuration only.
- A ≥ 1 criterion (union) would have admitted every feature picked up by any single configuration, inflating the predictor set with selections potentially driven by the specific behaviour of one particular method-model pair.
- A substantially stricter criterion (e.g. selection across a majority of the 24 configurations) would have been excessively conservative given the heterogeneity of the pipeline: univariate Cox and permutation importance produce quantitative rankings, whereas Boruta produces a binary decision, and some models naturally retain fewer features than others. A strict criterion would therefore structurally penalise the more parsimonious configurations.

It should be acknowledged that the $\geq 2/24$ threshold was relatively permissive: a feature may meet it through selection by a single method applied to two different models, rather than through agreement across methods. This is further discussed in Chapter 4.5. The criterion is best understood as a filter against single-configuration artefacts rather than as a strong multi-method consensus.

2.2.4 Model selection

Following the feature selection step, the consensus subset of predictors identified in Section 2.2.3 was carried forward to the modelling stage. The role of the model selection step described here was to fix the hyperparameters of each candidate estimator using only the training set; no model was eliminated at this stage, and the test set, strictly isolated since the initial train/test split, remained untouched throughout. The eight calibrated models were then all be carried forward to the final evaluation (Section 2.2.5).

Candidate models Eight estimators were calibrated, chosen so as to span the five families of survival models introduced in Section 1.5.3. The panel included the standard Cox proportional hazards model, its three penalised variants (Cox-Ridge, Cox-LASSO and Cox-ElasticNet, all solved by coordinate descent), the Random Survival Forest, two Gradient Boosting Survival variants (one with regression trees as weak learners, one componentwise), and the Survival Support Vector Machine in its fast formulation with a Gaussian Radial Basis Function (RBF) kernel. All estimators were instantiated through the `scikit-survival` library [95].

The rationale for this panel was twofold. First, it covered the full range of structural assumptions reviewed in the theoretical chapter (see Section 1.5.3): linear vs. non-linear dependence on the covariates, proportional vs. unconstrained hazards, hazard-based vs. ranking-based estimation. Carrying eight models through hyperparameter tuning therefore avoided committing prematurely to a single hypothesis class and allowed the final evaluation to compare what each modelling philosophy can extract from the same predictor set. Second, within the Cox family, retaining all three penalised variants alongside the unpenalised baseline let us probe separately the effect of pure shrinkage (Ridge), pure sparsity (LASSO), and their combination (ElasticNet), which behave very differently in the presence of correlated radiomic features.

Optimisation procedure For each model, an exhaustive grid was defined so as to cover a relevant range around the library’s default values while remaining computationally tractable. The values explored are summarised in Table 2.6. For each combination of each model, performance was estimated by 5-fold cross-validation on the training set with random shuffling (seed 123). Within each fold, predictors were standardised using a `StandardScaler` fitted on the training portion of the fold alone and then applied to the validation portion; this precaution prevented any information leakage between folds. The optimisation criterion was the C-index, averaged over the five folds. For each model, the hyperparameter combination that maximised this average C-index were retained.

The C-index was preferred to the other two metrics introduced in Section 1.5.4 for two reasons. First, it is the only metric that yielded a single, directly comparable scalar for every model in the panel. The Brier score requires absolute survival probabilities $\hat{S}(t | X)$ and is therefore undefined for the SSVM, which returns only a relative ranking score. The Cumulative/Dynamic AUC is in principle computable for all eight estimators, including the SSVM, as it requires only a risk score, but it depends on the choice of evaluation times t (see Section 2.2.5) and consequently produces a curve rather than a single value, which complicates direct cross-model comparison at the selection stage. Second, it is the standard metric in the survival-radiomics literature, which keeps the present results comparable with previously published pipelines.

2.2.5 Final evaluation

At the end of the model selection step (Section 2.2.4), each of the eight candidate models were associated with a single, fully specified hyperparameter configuration. Every model was then refitted once on the entire training set using these optimal hyperparameters and scored on the held-out test set, which has remained strictly untouched since the initial train/test split. This single, non-iterated evaluation provided an unbiased estimate of the generalisation performance of each model, complementary to the cross-validated performance obtained

Table 2.6: Hyperparameter grids explored using exhaustive search for each model. The parameter names correspond to those in the `scikit-survival` API.

Model	Hyperparameter	Values explored	<i>n</i>
CoxPH	alpha	{0.0001; 0.001; 0.01; 0.1; 1.0; 10.0}	6
Cox-Ridge	l1_ratio alphas	{0.001} {[0.001]; [0.01]; [0.05]; [0.1]; [0.5]; [1.0]; [5.0]}	7
Cox-LASSO	l1_ratio alphas	{1.0} {[0.005]; [0.01]; [0.03]; [0.05]; [0.1]; [0.3]}	6
Cox-ElasticNet	l1_ratio alphas	{0.1; 0.3; 0.5; 0.7; 0.9} {[0.01]; [0.03]; [0.1]; [0.3]; [1.0]}	25
RSF	n_estimators max_features min_samples_leaf min_samples_split max_depth max_samples	{300; 500; 1000} {sqrt; 0.5} {10; 15; 20} {10; 20} {4; 6; 8} {0.7; 1.0}	216
GBSM	n_estimators learning_rate max_depth subsample min_samples_leaf dropout_rate	{100; 200; 400} {0.01; 0.05; 0.1} {1; 2; 3} {0.5; 0.7} {10; 20} {0.0; 0.1}	216
GBSM CompWise	n_estimators learning_rate subsample dropout_rate	{100; 200; 400; 600} {0.01; 0.05; 0.1} {0.5; 0.7; 1.0} {0.0; 0.1; 0.02}	108
SSVM (RBF)	alpha gamma rank_ratio	{0.01; 0.1; 1.0; 10.0} {0.01; 0.05; 0.1; 0.5} {0.5; 0.8; 1.0}	48

during model selection. The eight models were scored along the three complementary axes introduced in Section 1.5.4, global discrimination, time-resolved discrimination and calibration, and the resulting estimates were accompanied by 95% confidence intervals quantified by non-parametric bootstrap.

Data preparation Before evaluation, the test set was restricted to the consensus subset of features retained in Section 2.2.3. A z-score standardisation was then applied: the per-feature mean and standard deviation were computed exclusively on the training set via `StandardScaler` (`scikit-learn`) and applied unchanged to the test set, to prevent any information leakage from the test data into the predictors.

The Cumulative/Dynamic AUC and the time-dependent Brier score (and therefore the IBS) further required an evaluation grid of time points, the choice of which was constrained by the behaviour of the IPCW estimators at the extremes of the follow-up distribution: at early times, the small number of observed events makes the AUC very noisy; at late

times, the depletion of the at-risk set in the training cohort causes the inverse censoring weights to explode, destabilising both estimators [50]. Following the recommendation of the `scikit-survival` documentation [50], the evaluation grid was set to 100 equidistant points between the 10th and 80th percentiles of the observed follow-up times in the training set. These bounds were deliberately conservative: the lower bound ensured that enough events had already accrued to stabilise the $\text{AUC}(t)$, while the upper bound stayed clear of the region in which the KM censoring estimator became uninformative.

Proportional-hazards assumption Because five of the eight candidate models belonged to the Cox family, the Proportional Hazards (PH) assumption underlying these models was verified on the final multivariate Cox model fitted on the training set with the consensus features. A Cox model was fitted with the `lifelines` library [96] using a light L_2 penalisation ($\text{penalizer} = 0.01$) for numerical stability, and the PH assumption was tested per feature using the scaled Schoenfeld residuals test with a rank transformation of the event times. A family-wise corrected global p -value was obtained by Bonferroni correction of the per-feature p -values.

Computed metrics The three metrics introduced in Section 1.5.4, the concordance index (C-index), the Cumulative/Dynamic AUC and the time-dependent Brier score, were computed for each model on the test set. The C-index and the Cumulative/Dynamic AUC were reported for all eight models, since both metrics only require a relative risk score. The IBS, in contrast, requires an absolute predicted survival probability $\hat{S}(t | X)$ at every grid point and was therefore undefined for the SSVM, which returns only a ranking score (Section 1.5.3); the SSVM was consequently excluded from this metric alone.

Uncertainty quantification by bootstrap For each model and each metric, a 95% confidence interval was estimated by non-parametric bootstrap of the test set. $B = 1000$ resamples were drawn with replacement from the test set, each of the same size as the original. For each bootstrap replicate, the predictions of the model, trained once on the full training set, and never refitted, were subsetting according to the resampled indices, and the three metrics were recomputed on this resampled subset. The 95% confidence interval was obtained by the percentile method, that is, by the 2.5% and 97.5% quantiles of the bootstrap distribution of each metric. The point estimate reported alongside each interval corresponds to the metric computed on the full, non-resampled test set.

Risk-group stratification A fourth axis of comparison was added to the three metric-based ones described above, intended to assess whether each model could translate its continuous risk score into a clinically interpretable dichotomy. For each model, the median of the risk scores predicted on the training set was used as a stratification threshold, and was then applied unchanged to the test set: test patients with a predicted risk at or below this threshold were assigned to the low-risk group, the remaining patients to the high-risk group. Computing the threshold on the training set, rather than on the test set itself, followed the same train/test separation principle maintained throughout the previous steps and prevented the cut-off from depending on the data on which it was evaluated; as a consequence, the two groups were not constrained to be of equal size at test time. The separation between the two

resulting Kaplan-Meier (Section 1.5.2) curves was then tested with a log-rank test and was summarised along two complementary axes, all implemented with the `lifelines` library [96].

- First, a log-rank test produced a p-value quantifying the statistical evidence against the null hypothesis of identical underlying survival distributions in the two groups.
- Second, the survival experience of each group was summarised by its RMST at horizon $\tau = 24$ months (≈ 730 days), introduced in Section 1.5.4. Per-group RMSTs were estimated as the area under the corresponding KM curve up to τ via the `lifelines` library [96], and their difference Δ -RMST was reported as the average number of event-free days gained by a low-risk patient over a high-risk patient within the $[0, \tau]$ window.

Chapter 3

Results

3.1 Feature selection

Feature reduction across stages

The feature selection pipeline produced a sharp, multi-stage reduction of the initial feature space (Table 3.1). The initial extraction yielded 10,204 features per patient. The unsupervised Spearman filter ($|\rho| > 0.75$) reduced this to 1,133 features, a single-step reduction of $-9,071$ features, that is approximately 88.9% of the original space. The union of features retained by at least one of the 24 supervised method $\times$ model combinations brought the set down to 20 features, a further reduction of $-1,113$. Finally, the consensus criterion (≥ 2 of 24 combinations) yielded 9 features, a marginal further reduction of only -11 .

Table 3.1: Number of features retained at each stage of the selection pipeline, with the reduction achieved at each step.

Stage	Features retained	Reduction
Initial extraction (per patient)	10,204	—
After unsupervised Spearman filter ($ \rho > 0.75$)	1,133	$-9,071$
Selected by ≥ 1 combinations (union)	20	$-1,113$
Selected by ≥ 2 combinations (consensus)	9	-11

Origin of the redundancy removed by the unsupervised filter To characterise which extraction axes generated the redundancy removed at the first stage of the pipeline, the 9,071 feature pairs flagged by the Spearman filter ($|\rho| > 0.75$) were decomposed along the four metadata axes used during extraction: MRI modality, tumour sub-region, radiomic class and image type. Each matrix was symmetric; diagonal entries quantified within-category redundancy, off-diagonal entries quantified between-category redundancy. All percentages were expressed with respect to the total number of correlated pairs along the corresponding axis.

By tumour sub-region (Figure 3.1), redundancy was overwhelmingly within sub-region: 27.7% of all correlated pairs involved two ED features, 29.1% two ET features and 31.1% two NCR features, giving a diagonal sum of 87.9%. Between-region redundancy was comparatively marginal: ED $\times$ ET and ET $\times$ NCR accounted for 2.9% and 2.0% respectively, and ED $\times$ NCR for only 1.2%. The three sub-regions therefore contributed roughly equally to the total redundancy, with NCR slightly above ED and ET.

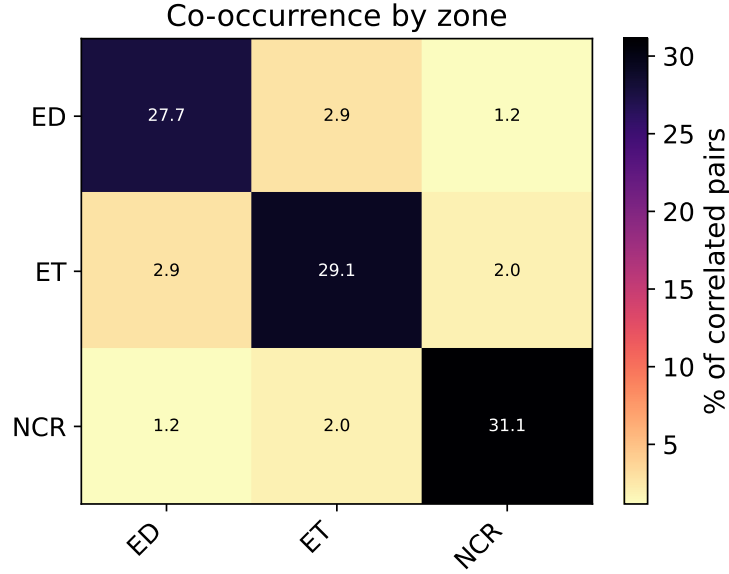


Figure 3.1: Co-occurrence matrix of correlated feature pairs ($|\rho| > 0.75$) decomposed by tumour sub-region. Percentages are expressed with respect to the total number of correlated pairs.

By MRI modality (Figure 3.2), redundancy was in contrast distributed almost uniformly across the 16 cells, with values ranging from 5.3% (FLAIR $\times$ FLAIR) to 6.9% (T1 $\times$ T1ce). Within-modality cells were not systematically higher than between-modality cells. The strongest between-modality block was T1 $\times$ T1ce (6.9%), followed by FLAIR $\times$ T1ce (6.7%), FLAIR $\times$ T2 (6.7%) and T1ce $\times$ T2 (6.4%). 4,937 pairs were excluded from this matrix because they involved shape features (which are modality-independent) or age; the values shown therefore refer to the modality-resolved subset only.

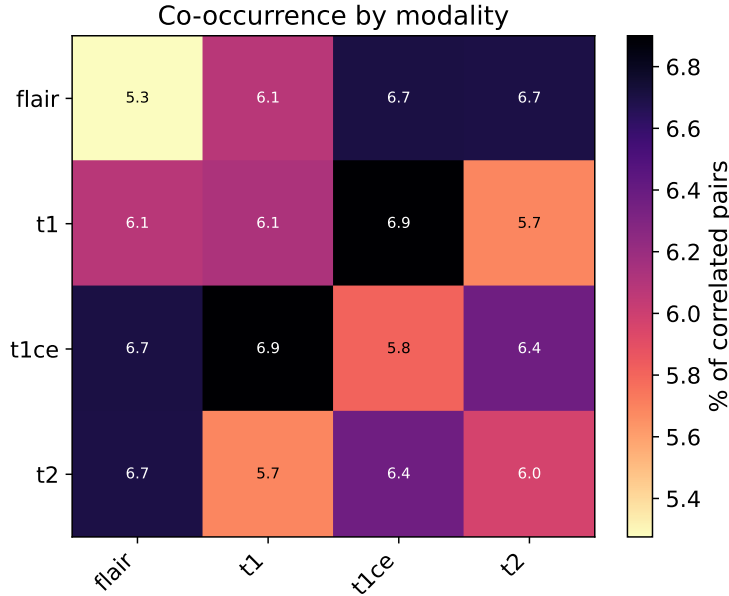


Figure 3.2: Co-occurrence matrix of correlated feature pairs decomposed by MRI modality. Pairs involving shape features or unknown entries are excluded.

By radiomic class (Figure 3.3), redundancy was markedly concentrated in two cells: GLDM $\times$ GLRLM

(17.9%) and GLRLM $\times$ GLRLM (15.7%), followed by GLDM $\times$ GLDM (7.0%) and GLCM $\times$ GLCM (4.9%). The GLSZM and NGTDM classes contributed small percentages ($\leq 3.6\%$ on the diagonal), and the shape row was almost empty ($\leq 0.3\%$ across all entries). The class axis was therefore the most uneven of the four: a single class pair (GLDM $\times$ GLRLM) accounted for almost one fifth of all correlated pairs removed at this stage.

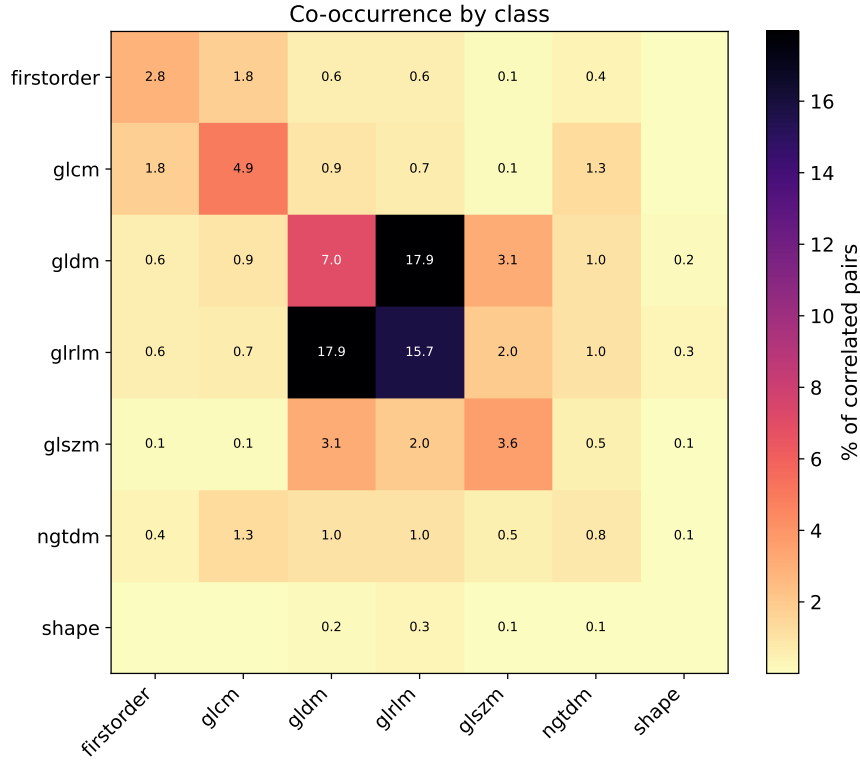


Figure 3.3: Co-occurrence matrix of correlated feature pairs decomposed by radiomic class.

By image type (Figure 3.4), the distribution was the most diffuse of the four matrices: all values lay between 0.5% and 2.1%. The diagonal entries (within the same wavelet sub-band or within the original image) were not systematically larger than the off-diagonals, and the highest entries were in fact between-sub-band cells (wavelet-HHH $\times$ wavelet-HLH and wavelet-HHH $\times$ wavelet-LHH, both 2.1%). The original image contributed most strongly to redundancy when paired with the wavelet-LLL sub-band (original $\times$ wavelet-LLL = 1.7%), which is the wavelet decomposition closest to the original signal. The two extreme sub-bands LLL and HHH otherwise generated comparatively little redundancy with the other wavelet sub-bands.

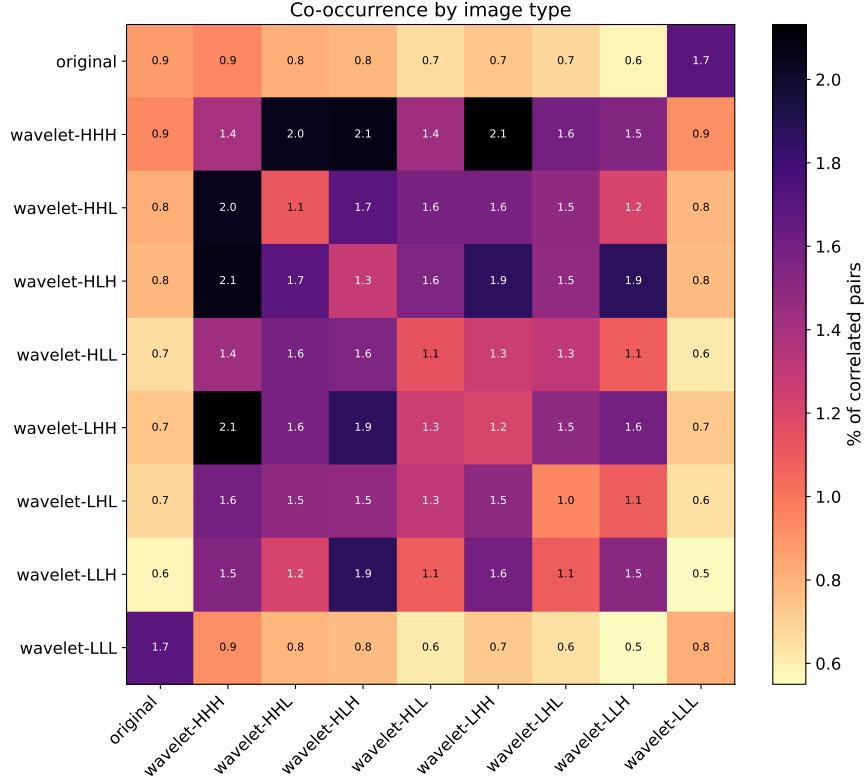


Figure 3.4: Co-occurrence matrix of correlated feature pairs decomposed by image type (original image and wavelet sub-bands).

Distribution of univariate Cox ranking scores ($n = 1,133$)

The ranking score $|C\text{-index} - 0.5|$ (Figure 3.5) ranged from ≈ 0 to 0.148 and was strongly right-skewed. Its mode lay at $|C\text{-index} - 0.5| \approx 0.005\text{--}0.010$ (the tallest bin holding on the order of 90 features), followed by a rapid decay beyond 0.04. The bulk of the distribution sat at low scores: 1,017 of the 1,133 features (89.8%) scored at or below 0.05, leaving only 116 (10.2%) above it. Of these, 114 fell in $(0.05, 0.10]$, 2 in $(0.10, 0.15]$, and none exceeded 0.15. The three highest-scoring features were:

- Age (C-index = 0.648, ranking score 0.148);
- `t1_original_glszm_GrayLevelNonUniformity_ET` (C-index = 0.605, ranking score 0.105);
- `t1_original_ngtdm_Strength_NCR` (C-index = 0.595, ranking score 0.095).

Age sat in isolation at the right edge of the distribution: its ranking score was 41% higher than that of the second-ranked feature (0.148 vs. 0.105, a C-index gap of 0.043), and no other feature approached its discriminative power. The full top-thirty is reported in Table A.2; two of these thirty features were *protective* (C-index = 0.425 and 0.428, i.e. ranking scores 0.075 and 0.072), meaning they ordered patients in the reverse direction but still carried usable prognostic information once their sign was recovered by a model.

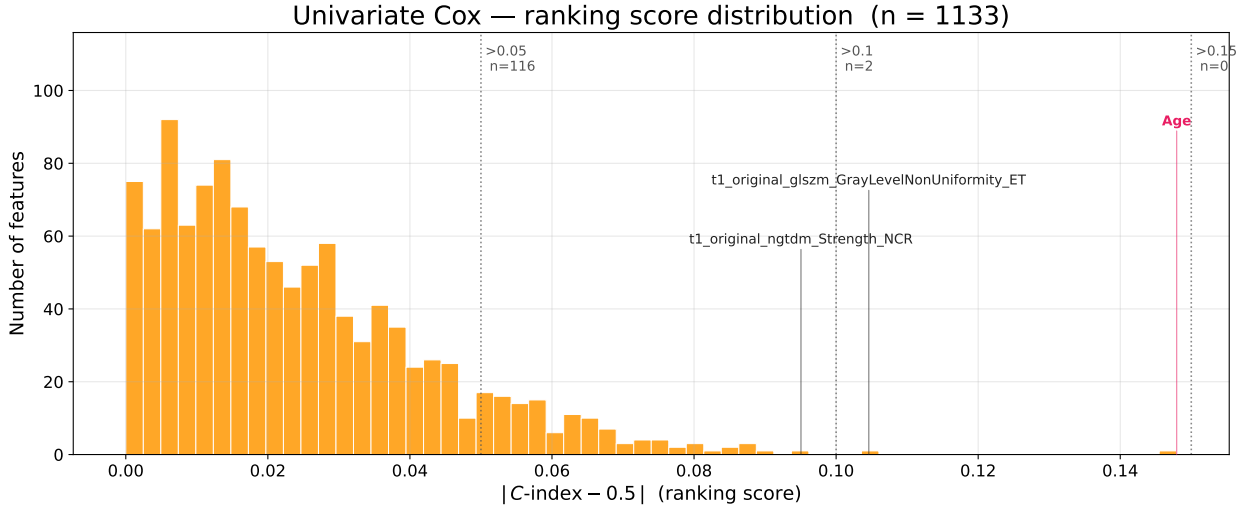


Figure 3.5: Distribution of the univariate Cox ranking scores ($|C\text{-index} - 0.5|$) on the 1,133 features remaining after the Spearman filter. The three highest-scoring features are annotated; vertical dotted lines mark the > 0.05 , > 0.10 and > 0.15 thresholds.

Distribution of permutation importance scores ($n = 1,133$)

The permutation-importance distribution (Figure 3.6) was far more concentrated than the Cox one. Scores ranged from ≈ -0.0005 to 0.0068 , and the modal bin centred on $\Delta C\text{-index} \approx 0$ alone gathered ≈ 625 features ($\approx 55\%$ of the total), a small fraction of which took slightly negative values (down to ≈ -0.0005). The informative tail was short: 1,099 features (97.0%) scored at or below 0.001 , so only 34 (3.0%) rose above it. Within that tail, 31 features fell in $(0.001, 0.002]$, just 2 in $(0.002, 0.005]$, and a single one exceeded 0.005 . The three highest-ranked features were:

- **Age** ($\Delta C\text{-index} = 0.0068$);
- **t1_wavelet-LHL_firstorder_Mean_ED** ($\Delta C\text{-index} = 0.0028$);
- **flair_original_shape_Elongation_ET** ($\Delta C\text{-index} = 0.0020$).

Age was again isolated at the right edge, and far more so than under the Cox score: its importance was $2.4\times$ that of the second-ranked feature (0.0068 vs. 0.0028 , a gap of 0.0040), a larger relative gap than for the Cox ranking, where **age** led the runner-up by a factor of only 1.4. Beyond **Age**, the tail was almost empty: the $(0.002, 0.005]$ band held just 2 features, so the gap between **age** and the near-zero bulk was bridged by virtually no intermediate scores. The full top-thirty of permutation importance is reported in Table A.3.

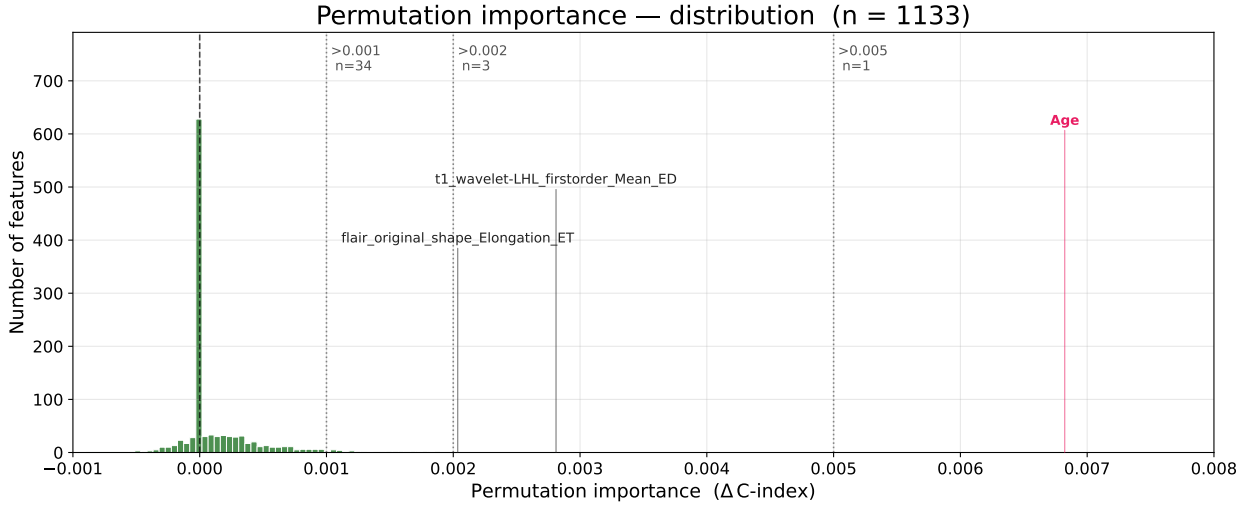


Figure 3.6: Distribution of the permutation-importance scores (Δ C-index) computed inside an RSF fitted on the 1,133 features. The three top-ranked features are annotated.

Concordance between univariate Cox and permutation importance

The scatter plot of Cox score against permutation score for the 1,133 features (Figure 3.7) showed no consensus feature in the lower-left corner. **Age** appeared alone in the upper-right corner, well separated from the eight remaining consensus features on both axes. The other consensus features distributed into two groups:

- a "low Cox, moderate permutation" group ($|C\text{-index} - 0.5| \approx 0.04$, $\Delta C\text{-index} \approx 0.002\text{--}0.003$):
 - t1_wavelet-LHL_firstorder_Mean_ED,
 - flair_original_shape_Elongation_ET,
 - flair_wavelet-LHH_glrml_HighGrayLevelRunEmphasis_ET;
- a "high Cox, low or negative permutation" group ($|C\text{-index} - 0.5| \approx 0.09\text{--}0.105$, $\Delta C\text{-index} \approx 0$ or slightly negative):
 - t1_original_ngtdm_Strength_NCR,
 - t1_original_glszm_GrayLevelNonUniformity_ET,
 - t2_original_ngtdm_Busyness_ED.

Two consensus features fell in an intermediate position with moderate scores on both axes:

- flair_original_glszm_GrayLevelNonUniformity_NCR
- t2_wavelet-LHL_firstorder_Kurtosis_ET

The grey cloud of non-consensus features was dense near (0, 0) and showed no clear global correlation between the two scores. Several non-consensus features individually outscored consensus features along one of the two axes.

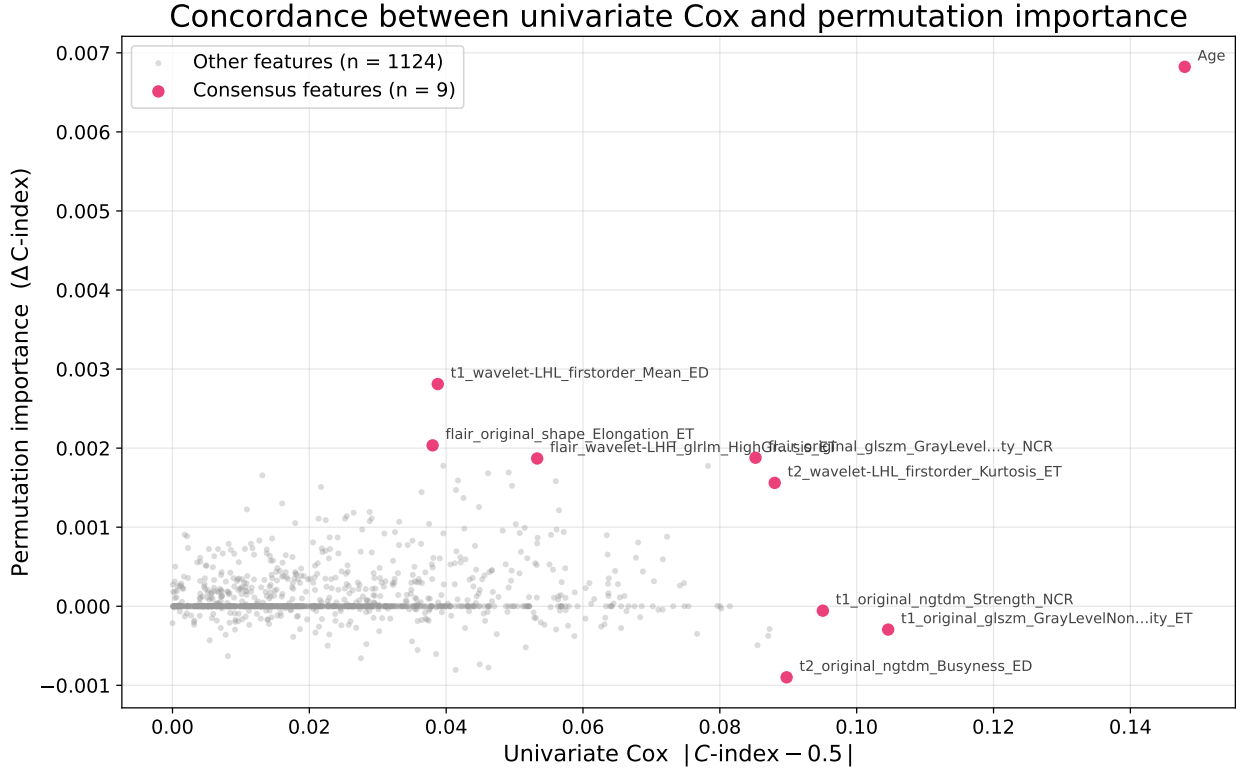


Figure 3.7: Permutation importance against the univariate Cox ranking score for the 1,133 features. The nine consensus features are highlighted and labelled.

Boruta

At convergence, **BorutaPy** classified 1 feature as confirmed, 0 as tentative, and 1,132 as rejected out of 1,133 input features. The single confirmed feature was **age**. Following the inclusive criterion defined in Section 2.2.3, both confirmed and tentative features were passed to the downstream models, which in this case reduced to **age** alone.

Effect of the 1-SD rule on the number of retained features

Figure 3.8 and Table 3.2 summarises the effect of the 1-SD rule for the 16 (method $\times$ model) combinations producing a ranking. The magnitude of the reduction varied considerably: Δk ranged from 0 (CoxPH $\times$ univariate Cox, Cox-Ridge $\times$ permutation) to 42 (SSVM $\times$ univariate Cox). Models with strong internal regularisation (Cox-LASSO, Cox-ElasticNet) tended to have small k_{max} and small k_{1-SD} , and the two values were close. Flexible non-regularised models (RSF, GBSM trees, SSVM) had large k_{max} (39–46) that the 1-SD rule reduced drastically to k_{1-SD} values between 1 and 14.

The associated C-index loss (ΔC -index) followed a similar pattern. The most aggressive reductions were accompanied by substantial losses: permutation $\times$ RSF (ΔC -index ≈ 0.030 , k from 41 to 4), univariate Cox $\times$ GBSM and permutation $\times$ GBSM (both ΔC -index ≈ 0.027). Conversely, several combinations showed no loss at all (CoxPH $\times$ univariate Cox, Cox-Ridge $\times$ permutation, both ΔC -index = 0). Across all combinations, k_{1-SD} concentrated around 4–5 features, while k_{max} spanned two orders of magnitude (1 to 46).

The plots showing the number of features kept for each combination are shown in Figures A.6 and A.7 in Appendix A.3.

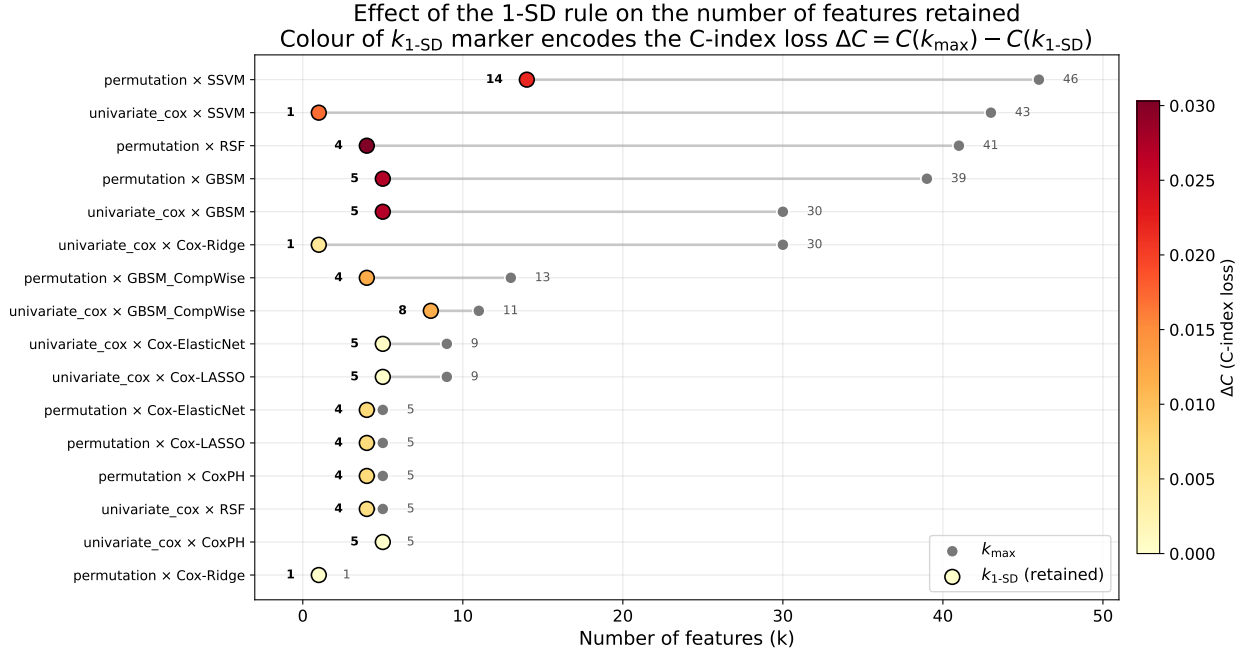


Figure 3.8: Effect of the one-standard-deviation rule on the number of features retained, for each of the 16 ranking-based method × model combinations. The colour of the k_{1-SD} marker encodes the C-index loss $\Delta C = C(k^*) - C(k_{1-SD})$. Boruta is omitted.

Table 3.2: Effect of the one-standard-deviation rule for each ranking-based method × model combination. k^* maximises the mean cross-validated C-index; k_{1-SD} is the smallest k whose mean C-index lies within one SE of $C(k^*)$. Boruta is omitted as it does not produce a ranking.

Method	Model	k^*	$C(k^*)$	Threshold	k_{1-SE}	$C(k_{1-SE})$	ΔC	Δk
Univariate Cox	CoxPH	5	0.692	0.682	5	0.692	0.0	0
	Cox-Ridge	30	0.657	0.642	1	0.652	0.005	29
	Cox-LASSO	9	0.692	0.682	5	0.692	0.000	4
	Cox-ElasticNet	9	0.693	0.682	5	0.692	0.001	4
	RSF	5	0.699	0.680	4	0.692	0.007	1
	GBSM (trees)	30	0.679	0.646	5	0.652	0.027	25
	GBSM componentwise	11	0.704	0.692	8	0.692	0.012	3
	SSVM (RBF)	43	0.651	0.616	1	0.634	0.022	42
Permutation	CoxPH	5	0.703	0.678	4	0.696	0.007	1
	Cox-Ridge	1	0.652	0.642	1	0.652	0.000	0
	Cox-LASSO	5	0.702	0.681	4	0.695	0.007	1
	Cox-ElasticNet	5	0.702	0.681	4	0.695	0.007	1
	RSF	41	0.706	0.661	4	0.676	0.030	37
	GBSM (trees)	39	0.682	0.646	5	0.655	0.027	34
	GBSM componentwise	13	0.709	0.683	4	0.697	0.012	9
	SSVM (RBF)	46	0.676	0.645	14	0.654	0.022	32

Selection frequency decomposed by method

Feature relevance was assessed over 24 method $\times$ model combinations. The two ranking methods (univariate Cox and permutation importance) produced, for each of the eight models, an ordered list from which the 1-SD rule set the number of retained features k , giving 16 ranking-based combinations. Boruta, which returned a binary decision rather than a ranking, needed no such cut-off and instead passed its confirmed-and-tentative subset unchanged to each of the eight models, giving the remaining 8 combinations. For every feature retained by at least one of the 24 combinations, we then counted how many retained it; the resulting frequency distribution is shown in Figure 3.9.

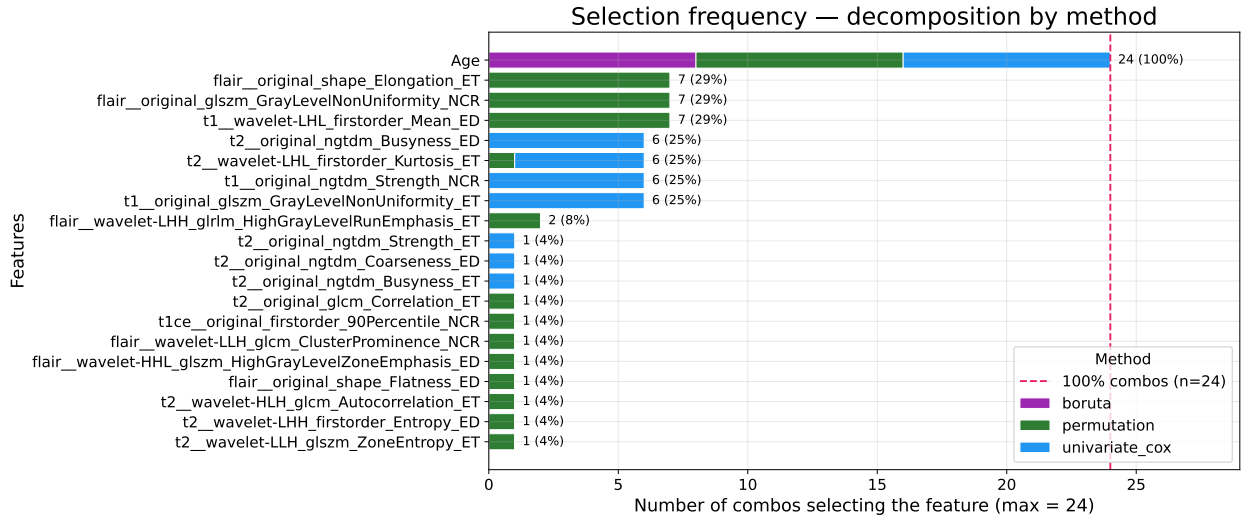


Figure 3.9: Selection frequency of each feature across the 24 method $\times$ model combinations, stacked by selection method. The dashed line marks the maximum (24/24).

The consensus feature set

Applying the consensus criterion (≥ 2 of 24 combinations) retained nine features, listed in Table 3.3 with their metadata and selection methods. Their distribution by selection method was polarised: **age** was the only feature retained by all three methods (U + P + B); one feature, **t2_wavelet-LHL_firstorder_Kurtosis_ET**, was retained by two methods (U + P); the remaining seven were each retained by a single method, four by permutation and three by univariate Cox.

Table 3.3: The nine consensus features (retained by ≥ 2 of the 24 method $\times$ model combinations), with their MRI modality, image filter, radiomic class, metric and tumour sub-region. B = Boruta, P = permutation importance, U = univariate Cox.

Modality	Filter	Class	Metric	Region	Methods	Freq./24
—	—	Clinical	Age	—	U + P + B	24
—	—	Shape 3D	Elongation	ET	P	7
FLAIR	original	GLSZM	GrayLevelNonUniformity	NCR	P	7
T1	wavelet-LHL	First Order	Mean	ED	P	7
T2	original	NGTDM	Busyness	ED	U	6
T2	wavelet-LHL	First Order	Kurtosis	ET	U + P	6
T1	original	NGTDM	Strength	NCR	U	6
T1	original	GLSZM	GrayLevelNonUniformity	ET	U	6
FLAIR	wavelet-LHH	GLRLM	HighGrayLevelRunEmphasis	ET	P	2

Composition of the consensus set The aggregate composition of the consensus set along the four metadata axes used during extraction is shown in Figure 3.10.

- By modality, three features were derived from T1 (33%), two from FLAIR and T2 each (22%), one was **age** (11%), and one was the shape feature (11%), reported here under a separate "Shape" category as it depended only on the segmentation mask. T1ce was entirely absent from the consensus, despite being one of the four extracted modalities.
- By radiomic class, three classes were tied at the top with two features each (22%): GLSZM, First Order, and NGTDM. Shape 3D and GLRLM each contributed one feature. GLCM and GLDM were completely absent from the consensus, although they together accounted for 38 of the 94 intensity- and texture-dependent features extracted per (modality, sub-region) combination.
- By tumour sub-region, ET dominated with four features (44%); NCR and ED each contributed two features (22%). All three sub-regions available in BraTS2020 were represented, but ET hosted twice as many features as either of the other two.
- By image filter, five features (56%) were computed on the original image. Only two of the eight wavelet sub-bands appeared: LHL with two features and LHH with one. The remaining six sub-bands (LLL, LLH, HLL, HLH, HHL, HHH) yielded no retained feature.

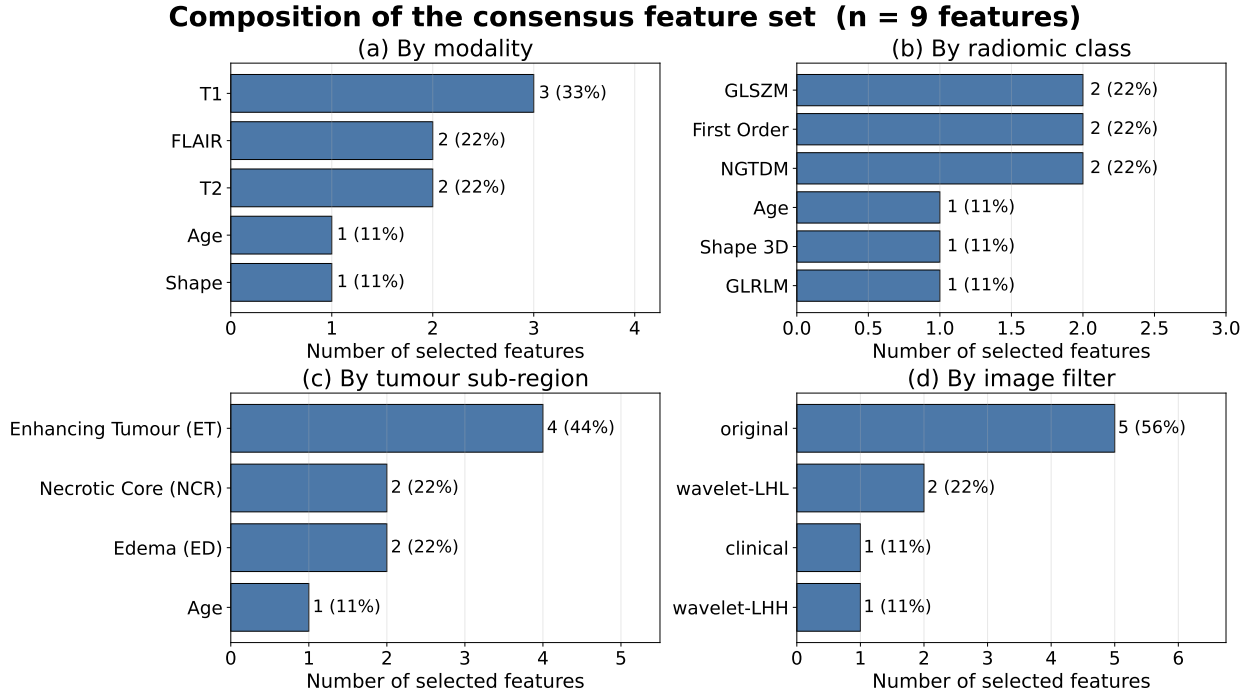


Figure 3.10: Composition of the nine consensus features along four axes: (a) MRI modality, (b) radiomic class, (c) tumour sub-region and (d) image filter. Counts are expressed in absolute numbers and percentages.

Joint distributions across pairs of metadata axes Three co-occurrence heatmaps describe how the consensus features distribute across pairs of metadata axes.

First, Figure 3.11 shows class $\times$ sub-region. ET was the most diverse sub-region in terms of classes (four: First Order, GLRLM, GLSZM, Shape 3D). ED and NCR each combined two classes (First Order + NGTDM for ED, GLSZM + NGTDM for NCR). NGTDM was the only retained class absent from ET, although ET was otherwise the most heavily represented sub-region.

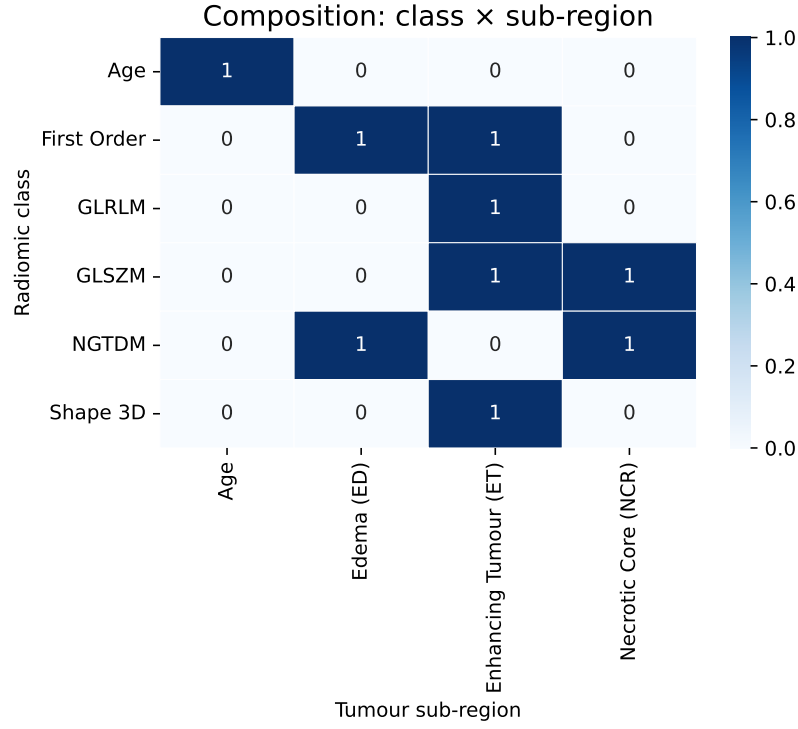


Figure 3.11: Joint distribution of the nine consensus features along radiomic class $\times$ tumour sub-region.

Then, Figure 3.12 shows modality $\times$ sub-region. T1 was the only modality with features in all three sub-regions (ED, ET, NCR). T2 was absent from NCR, FLAIR was absent from ED, and Shape was limited to ET. Conversely, ET was the only sub-region in which all four modality-or-Shape categories appeared simultaneously. T1ce remained absent in every cell.

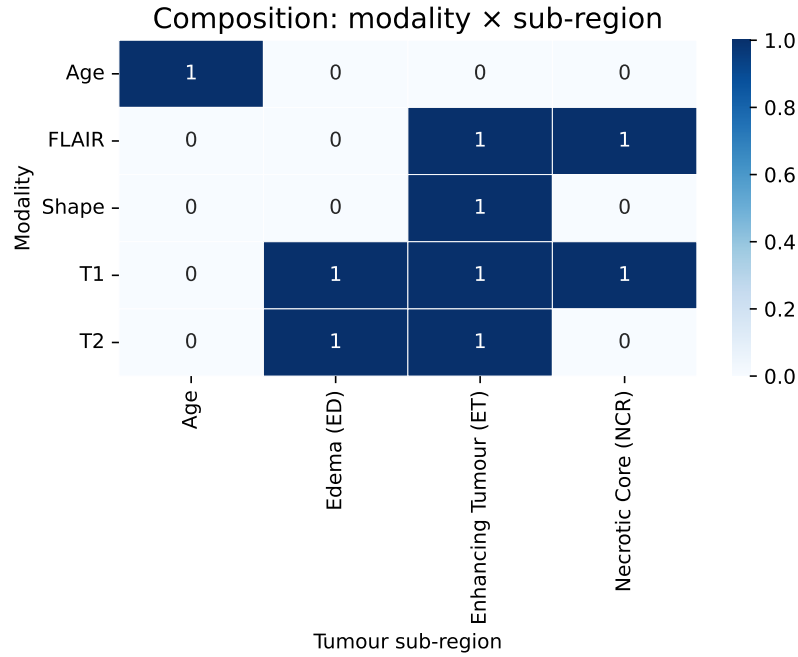


Figure 3.12: Joint distribution of the nine consensus features along MRI modality $\times$ tumour sub-region.

Finally, Figure 3.13 shows class $\times$ modality. Three of the five classes (First Order, GLSZM, NGTDM) spanned two modalities; GLRLM was confined to FLAIR and Shape 3D to the Shape category. T1 covered the largest number of classes (three), followed by FLAIR and T2 (two each). Every retained texture class was represented in T1, with the single exception of GLRLM.



Figure 3.13: Joint distribution of the nine consensus features along radiomic class $\times$ MRI modality.

Internal correlation structure of the consensus set The Spearman correlation matrix computed on the nine consensus features (Figure 3.14) shows that all pairwise absolute correlations remained below 0.55, well under the $|\rho| = 0.75$ threshold of the upstream Spearman filter. Only three pairs exceeded $|\rho| = 0.30$:

- $\rho = +0.53$ between `flair_original_glszm_GrayLevelNonUniformity_NCR` and `t1_original_glszm_GrayLevelNonUniformity_ET`, the same metric in two different sub-regions and modalities, and the strongest positive association in the matrix;
- $\rho = -0.37$ between `t1_original_ngtdm_Strength_NCR` and `t1_original_glszm_GrayLevelNonUniformity_ET`;
- $\rho = -0.34$ between `t1_original_ngtdm_Strength_NCR` and `t1_wavelet-LHL_firstorder_Mean_ED`.

Age showed uniformly weak associations with all radiomic features ($|\rho| \leq 0.24$, the largest being with `t2_original_ngtdm_Busyness_ED`). `flair_wavelet-LHH_glrlm_HighGrayLevelRunEmphasis_ET` had $|\rho| \leq 0.05$ with every other consensus feature and was the most isolated element of the set. Among the moderate-magnitude correlations, negative values predominated; `t1_original_ngtdm_Strength_NCR` in particular showed weak-to-moderate negative correlations with seven of the eight other features.

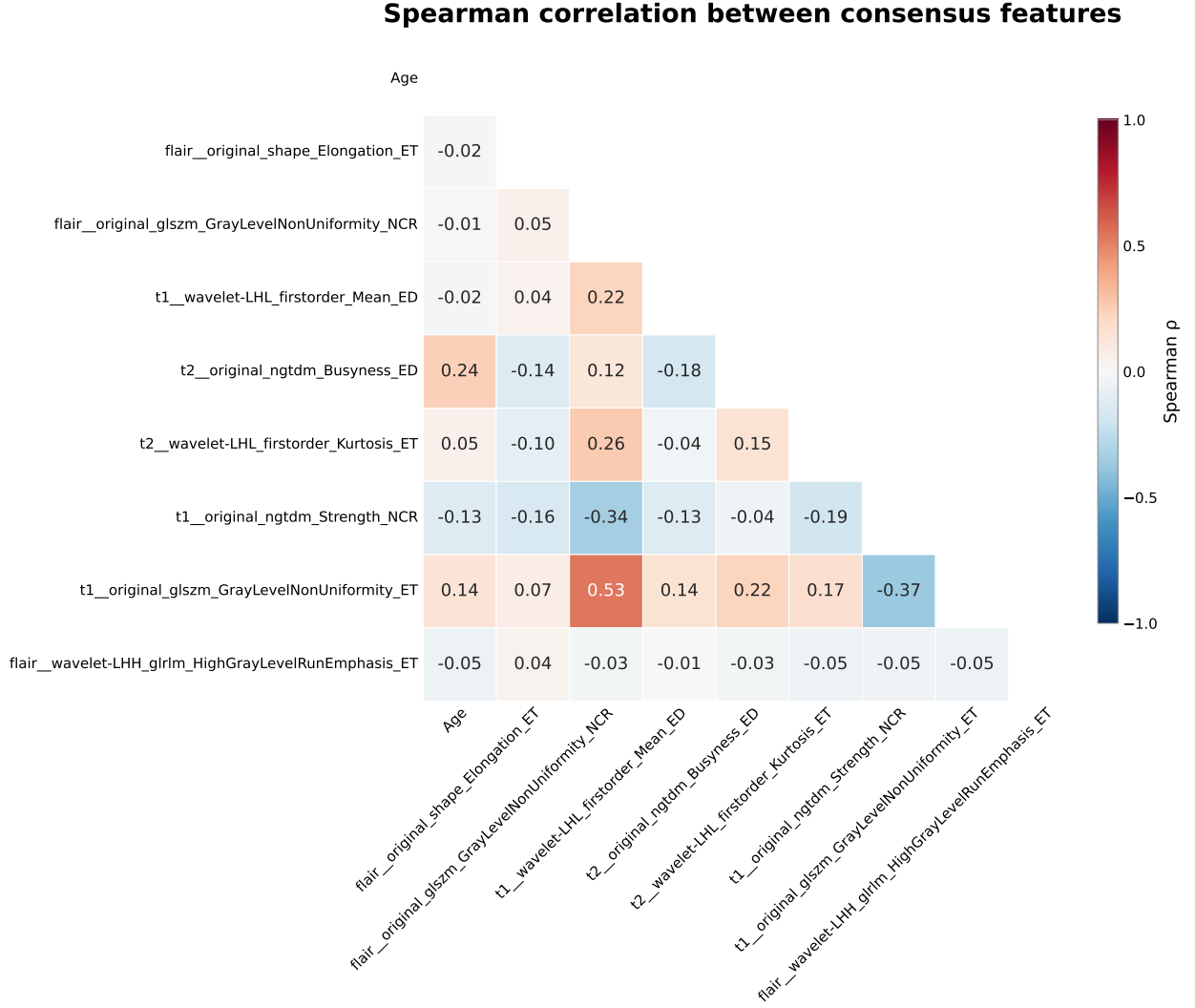


Figure 3.14: Pairwise Spearman rank correlations between the nine consensus features, computed on the training set ($n = 187$). Only the lower triangle is displayed.

3.2 Survival modelling

Before comparing the eight candidate models, the Proportional Hazards (PH) assumption was checked on the final multivariate Cox model fitted on the training set ($n = 187$) with the nine consensus features. The fitted model reached a training concordance of 0.703 (log-likelihood -754.34 , partial AIC 1526.69). The scaled Schoenfeld residuals test (Table 3.4) flagged a single feature as violating PH at the 0.05 level, **age** ($p = 0.041$), while the remaining eight features showed no evidence of non-proportionality ($p \geq 0.20$). After Bonferroni correction for the nine simultaneous tests, the global hypothesis of proportional hazards was not rejected ($p_{\text{global}} = 0.37$). The PH assumption was therefore considered globally satisfied, which supported the use of the Cox-based models in the comparison that follows; the borderline, uncorrected violation for **age** was interpreted with mild caution.

Table 3.4: Schoenfeld residuals test of the PH assumption for the final multivariate Cox model (nine consensus features, training set, $n = 187$). Features are ordered by increasing p -value. The global p -value is the Bonferroni-corrected minimum.

Feature	Test stat.	p -value	Violates PH?
Age	4.160	0.0414	yes
t1_original_ngtdm_Strength_NCR	1.611	0.2044	no
flair_original_shape_Elongation_ET	1.418	0.2338	no
t2_wavelet-LHL_firstorder_Kurtosis_ET	0.979	0.3225	no
flair_wavelet-LHH_glrml_HighGrayLevelRunEmphasis_ET	0.886	0.3465	no
t2_original_ngtdm_Busyness_ED	0.412	0.5209	no
t1_original_glszm_GrayLevelNonUniformity_ET	0.239	0.6248	no
flair_original_glszm_GrayLevelNonUniformity_NCR	0.165	0.6844	no
t1_wavelet-LHL_firstorder_Mean_ED	0.050	0.8230	no
Global (Bonferroni)		0.3725	no

3.2.1 Cross validation performances per model

The eight candidate models were calibrated on the training set using 5-fold cross-validation on the consensus signature. The retained hyperparameter values are summarised in Table 3.5.

All models were evaluated by 5-fold cross-validation on the training set along the three metrics introduced in Section 1.5.4 (C-index, mean dynamic AUC, IBS). The aggregated values are reported in Table 3.6 and the per-fold distributions in Figure 3.15.

The cross-validation C-index spanned a narrow band of approximately 0.011, from 0.6829 (SSVM) to 0.6942 (GBSM). The mean dynamic AUC band was even narrower, from 0.7780 (SSVM) to 0.7832 (Cox-Ridge), an amplitude of approximately 0.005. The IBS band, for which SSVM was excluded by construction, ran from 0.1346 (Cox-ElasticNet) to 0.1428 (RSF), an amplitude of approximately 0.008.

The ranking was not consistent across metrics. GBSM was first on C-index but only fourth on AUC; Cox-Ridge was third on C-index but first on AUC; Cox-ElasticNet was seventh on C-index but first on IBS; RSF was second on C-index but last on IBS. The fold-to-fold variability also differed substantially across models: SSVM had by far the lowest standard deviation on both C-index (0.0110) and AUC (0.0064), with a noticeably tighter boxplot than the rest of the panel, whereas GBSM had the largest SD on every metric for which it was evaluated (0.0404 on C-index, 0.0540 on AUC, 0.0231 on IBS). All cross-validation AUC values remained above the random baseline of 0.5, and all IBS values lay well below the null-model reference of 0.25.

Table 3.5: Hyperparameter configurations retained for each model after the 5-fold cross-validated grid search on the training set. Parameter names follow the `scikit-survival` API.

Model	Hyperparameter	Selected value(s)
CoxPH	alpha	10.0
Cox-Ridge	l1_ratio	0.001
	alphas	[0.01]
Cox-LASSO	l1_ratio	1.0
	alphas	[0.03]
Cox-ElasticNet	l1_ratio	0.1
	alphas	[0.1]
RSF	n_estimators	1000
	max_features	sqrt
	min_samples_leaf	20
	min_samples_split	10
	max_depth	4
	max_samples	0.7
GBSM	n_estimators	400
	learning_rate	0.01
	max_depth	1
	subsample	0.5
	min_samples_leaf	20
	dropout_rate	0.0
GBSM componentwise	n_estimators	400
	learning_rate	0.05
	subsample	0.7
	dropout_rate	0.0
Survival SSVM (RBF)	alpha	0.1
	gamma	0.01
	rank_ratio	1.0

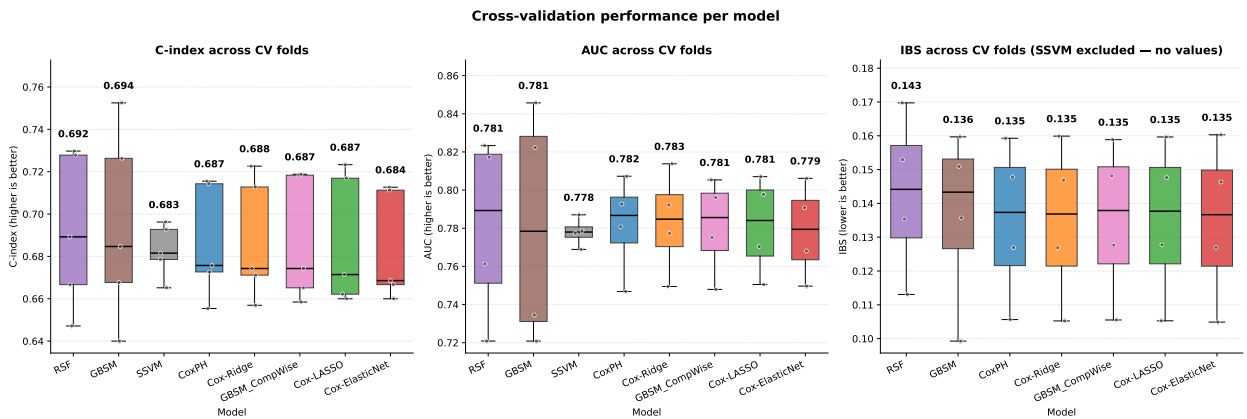
**Figure 3.15:** Per-fold distributions of the three performance metrics for the eight candidate models in 5-fold cross-validation on the training set: C-index, mean Cumulative/Dynamic AUC and Integrated Brier Score. The SSVM is excluded from the AUC and IBS panels. Values above each box are the fold-averaged means.

Table 3.6: Cross-validated performance of the eight candidate models on the training set, ranked by decreasing C-index. Values are means across the five validation folds, with standard deviations in parentheses. "—" indicates that the metric is undefined for the SSVM.

Model	C-index (SD)	Mean dynamic AUC (SD)	IBS (SD)
GBSM	0.6942 (0.0404)	0.7808 (0.0540)	0.1364 (0.0231)
RSF	0.6921 (0.0328)	0.7807 (0.0422)	0.1428 (0.0210)
Cox-Ridge	0.6876 (0.0255)	0.7832 (0.0234)	0.1347 (0.0207)
GBSM componentwise	0.6870 (0.0263)	0.7811 (0.0221)	0.1350 (0.0204)
Cox-LASSO	0.6868 (0.0276)	0.7814 (0.0224)	0.1351 (0.0206)
CoxPH	0.6867 (0.0240)	0.7819 (0.0223)	0.1349 (0.0205)
Cox-ElasticNet	0.6838 (0.0232)	0.7787 (0.0215)	0.1346 (0.0209)
SSVM	0.6829 (0.0110)	0.7780 (0.0064)	—

3.2.2 Test performance per model

Test-set performance with bootstrap confidence intervals

Each of the eight models was refitted once on the full training set with its calibrated hyperparameters and evaluated on the held-out test set. Point estimates and 95% non-parametric bootstrap confidence intervals ($n_{\text{boot}} = 1000$) are reported in Table 3.7; the corresponding bootstrap distributions are shown in Figure 3.16.

The test-set C-index ranged from 0.625 (Cox) to 0.671 (RSF), an amplitude of approximately 0.046, about four times wider than the cross-validation amplitude. The eight models separated into three groups:

- tree-based ensembles: RSF (0.671) and GBSM (0.662);
- SSVM (RBF) in an intermediate position (0.640);
- linear Cox-based models plus GBSM componentwise clustered between 0.625 and 0.631 (CoxPH 0.625, Cox-LASSO 0.626, Cox-Ridge 0.627, Cox-ElasticNet 0.629, GBSM componentwise 0.631).

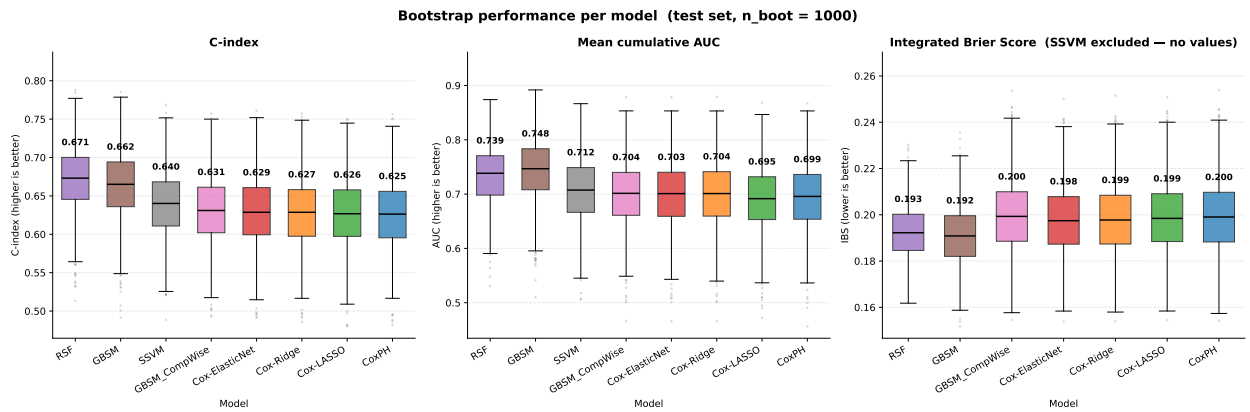
The same structure appeared on the mean AUC(t): GBSM was highest at 0.748, followed by RSF (0.739), SSVM (0.712), then the linear cluster between 0.695 and 0.704. On IBS, GBSM and RSF again sat at the lower (better) end (0.192 and 0.193 respectively), while the linear Cox-based models all clustered at 0.198–0.200.

The bootstrap distributions (Figure 3.16) confirmed the visual separation between the tree-based ensembles and the linear cluster. The inter-quartile ranges were visually similar across models, with no equivalent of the SSVM tightening observed in cross-validation. A small number of bootstrap outliers appeared in the lower tails of the C-index distributions for the linear models, close to 0.50.

The 95% bootstrap CIs were wide: each C-index Confidence Interval (CI) had a width of approximately 0.16, so every pair of CIs overlapped substantially. All C-index lower bounds remained above 0.50 except marginally, the smallest being 0.541 (Cox-LASSO).

Table 3.7: Test-set performance of the eight calibrated models. Brackets give the 95% bootstrap percentile confidence interval ($B = 1,000$). The SSVM is excluded from the IBS column.

Model	C-index [95 % CI]	Mean AUC(t) [95 % CI]	IBS [95 % CI]
CoxPH	0.625 [0.542–0.706]	0.699 [0.570–0.804]	0.200 [0.169–0.233]
Cox-Ridge	0.627 [0.544–0.707]	0.704 [0.578–0.810]	0.199 [0.168–0.231]
Cox-LASSO	0.626 [0.541–0.705]	0.695 [0.568–0.799]	0.199 [0.169–0.231]
Cox-ElasticNet	0.629 [0.547–0.710]	0.703 [0.576–0.810]	0.198 [0.168–0.230]
RSF	0.671 [0.584–0.750]	0.739 [0.626–0.823]	0.193 [0.172–0.216]
GBSM	0.662 [0.569–0.741]	0.748 [0.622–0.841]	0.192 [0.168–0.217]
GBSM componentwise	0.631 [0.548–0.712]	0.704 [0.579–0.807]	0.200 [0.169–0.233]
SSVM (RBF)	0.640 [0.555–0.714]	0.712 [0.580–0.815]	—

**Figure 3.16:** Bootstrap distributions of the three performance metrics for the eight calibrated models on the test set ($n = 47$, $B = 1,000$ non-parametric resamples). The SSVM is excluded from the IBS panel.

Temporal dynamics

The time-resolved dynamic AUC and Brier score curves are shown in Figures 3.17 and 3.18, between approximately 3 and 20.5 months of follow-up.

The dynamic AUC curves of all eight models followed a shared trajectory: initial values around 0.72–0.78 at month 3, a sharp dip around months 6.5–7, a peak around months 9–10, a progressive decline, a second sharp drop around month 15, and convergence to 0.55–0.65 at the end of the observation window. All curves remained above the random baseline of 0.5 throughout. RSF was the highest curve across most of the range, peaking near 0.83 around month 10 (mean AUC = 0.739). GBSM was the highest curve at very early times (months 3–6, around 0.79) and remained the second-highest overall (mean AUC = 0.739). SSVM sat between the tree-based ensembles and the linear cluster across most of the range (mean AUC = 0.712). The five linear Cox-based configurations (Cox, Ridge, LASSO, ElasticNet, CompWise) bunched within a band of approximately 0.01 and had mean AUCs between 0.695 (LASSO) and 0.704 (Ridge, CompWise).

The Brier score curves (SSVM excluded by construction) showed a complementary pattern: low values around month 5 (0.14–0.17), a sharp rise between months 6 and 8 to 0.20–0.21, a noisy plateau between months 8 and 15, a peak near month 13 (0.22–0.23), and a decline to

0.16–0.20 at the end of the window. All curves remained well below the null-model reference of 0.25. GBSM (IBS = 0.192) and RSF (IBS = 0.193) sat visibly below the linear cluster throughout the middle and late portions of the range, with a roughly constant offset of approximately 0.01–0.02 in absolute Brier score. The five linear-Cox-style models (CoxPH 0.200, Ridge 0.199, LASSO 0.199, Elastic 0.198, CompWise 0.200) tracked each other very closely and were largely indistinguishable visually. The overall shape of the trajectories was similar across all models; the tree-based ensembles were essentially shifted downward.

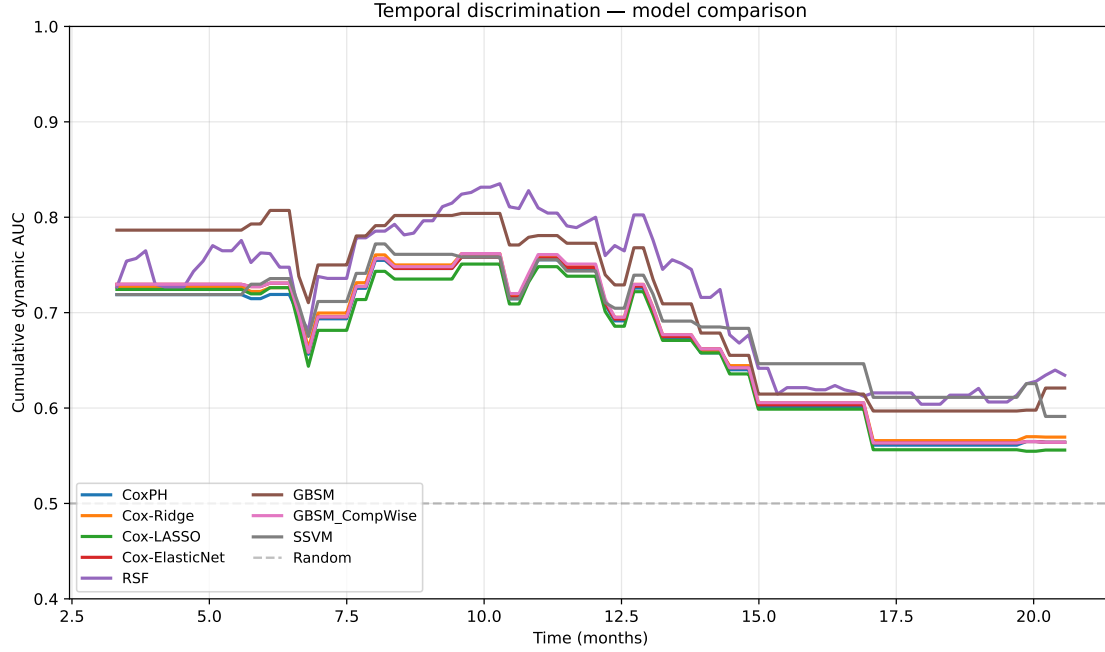


Figure 3.17: Cumulative/Dynamic AUC as a function of time for the eight calibrated models on the test set. The evaluation grid spans the 10th–80th percentile of the training follow-up; the dashed line at 0.5 marks the random-ranking baseline.

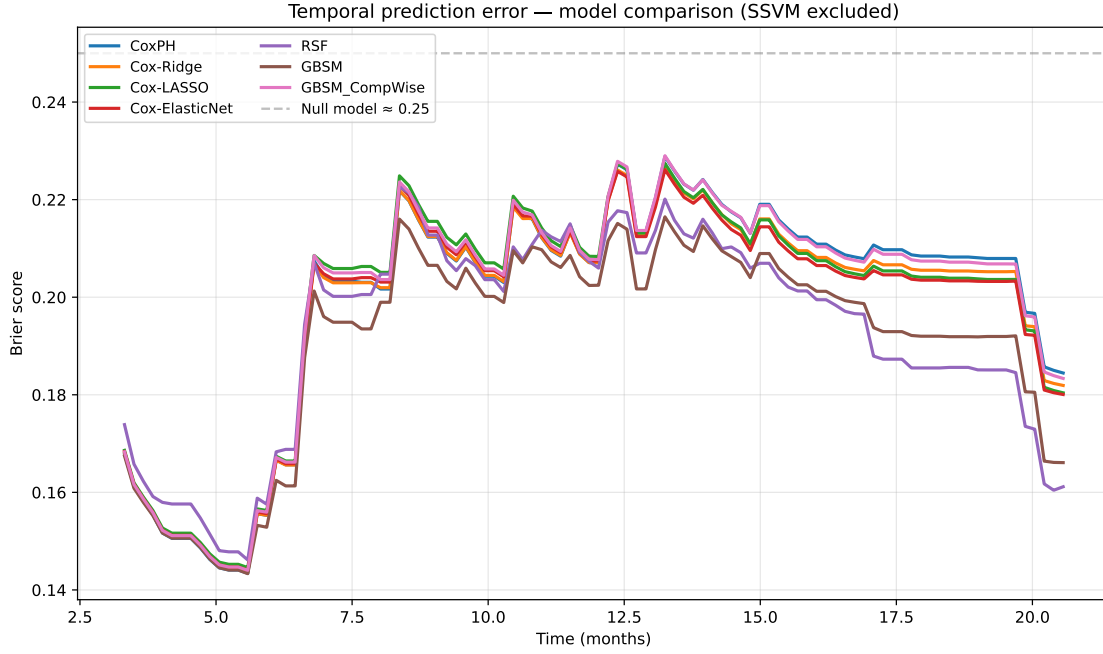


Figure 3.18: Brier score as a function of time for the seven calibrated models for which an absolute survival function is defined (SSVM excluded). The dashed line at 0.25 marks the upper bound of an uninformative model.

Risk-group stratification

KM curves and log-rank tests. The KM curves of the two predicted risk groups are shown for each of the eight models in Figure 3.19. Seven of the eight models produced a visual separation between the high- and low-risk groups across the follow-up window, with the high-risk curves dropping faster and reaching low survival probabilities first. The single exception was GBSM (tree-based), whose two curves overlapped substantially throughout follow-up and crossed around 750 days, indicating that the model failed to dichotomise patients into two distinct survival profiles.

The log-rank p-values reported in the panel titles confirmed this visual reading. Seven models reached conventional significance ($p < 0.05$), Cox-ElasticNet (1.18×10^{-2}), SSVM (1.31×10^{-2}), RSF (1.43×10^{-2}), Cox-LASSO (2.15×10^{-2}), CoxPH = Cox-Ridge (2.55×10^{-2}) and GBSM componentwise (4.69×10^{-2}), while GBSM (tree-based) did not (1.38×10^{-1}). None of the significant p-values fell below 0.01, so the entire significant panel sat in the 0.01–0.05 band; GBSM componentwise was borderline, only marginally below the 0.05 threshold.

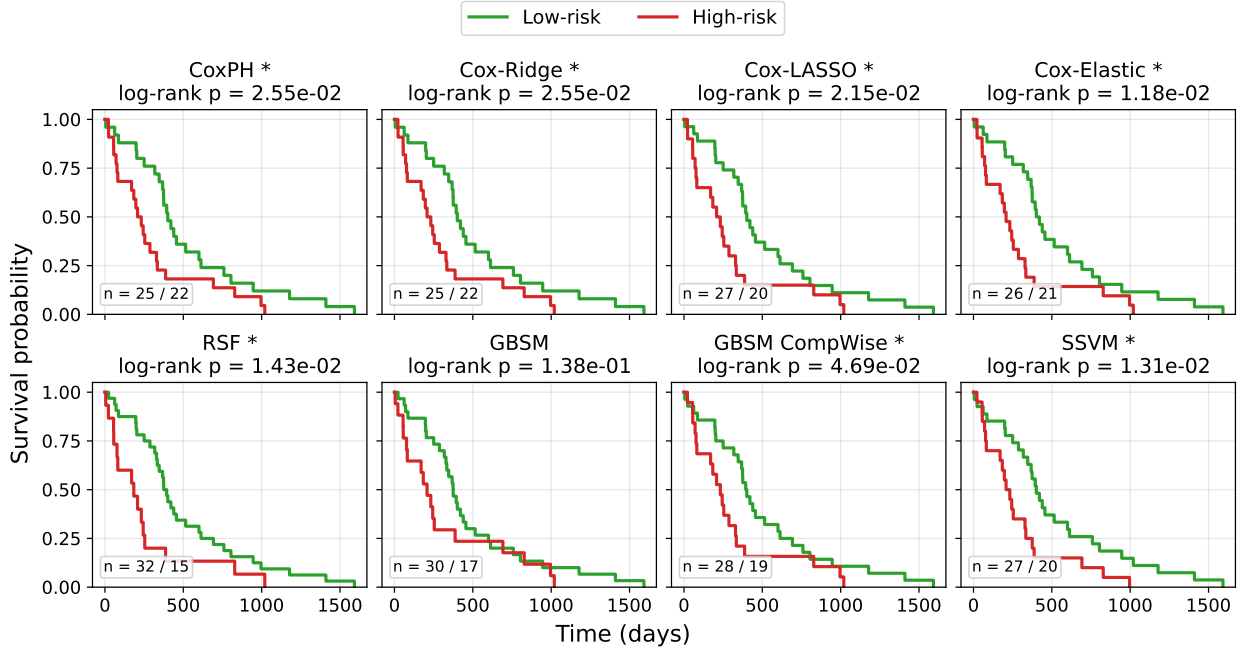


Figure 3.19: Kaplan–Meier survival curves of the predicted high- and low-risk groups for each of the eight candidate models on the test set. Risk groups are obtained by dichotomising the predicted risk score at its median; the log-rank p-value is reported in each panel title.

RMST At $\tau = 24$ months (Figure 3.20), all eight models produced a positive ΔRMST , indicating that the low-risk group accumulated more event-free time than the high-risk group over the two-year window for every model in the panel. The amplitude of ΔRMST across models was approximately 76 days, ranging from 116 days (GBSM tree-based) to 192 days (RSF). RSF, Cox-ElasticNet and Cox-LASSO occupied the top three positions, with ΔRMST s of 192, 185 and 173 days, respectively; CoxPH and Cox-Ridge followed at 155 days, ahead of SSVM (153 days) and GBSM componentwise (146 days); GBSM (tree-based) closed the ranking at 116 days. The model ordering produced by the ΔRMST broadly matched the ordering observed on the log-rank p-values of Figure 3.19, with GBSM (tree-based) remaining the only model that failed to separate the two risk groups by a clinically large margin.

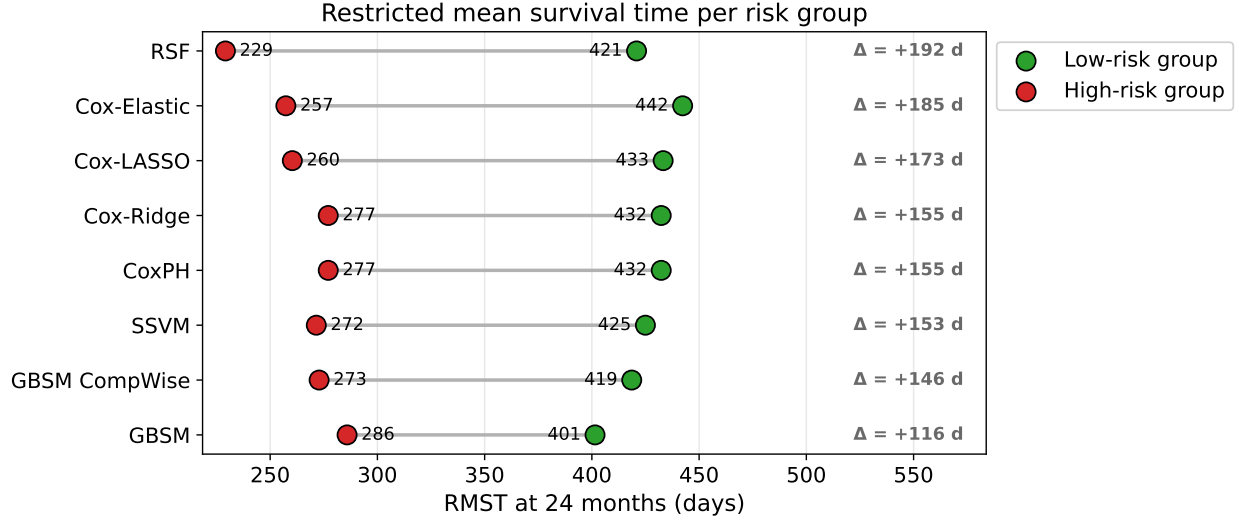


Figure 3.20: RMST at $\tau = 24$ months for the predicted high-risk (red) and low-risk (green) groups of each candidate model on the test set, ordered by decreasing between-group difference $\Delta \text{RMST} = \text{RMST}_{\text{low}} - \text{RMST}_{\text{high}}$ (in days, right).

3.3 Results of the selected model

Of the eight calibrated estimators compared on the held-out test set (Section 3.2.2, Table 3.7), the RSF was retained as the final model. This choice is discussed in Section 4.2. All the analyses reported in this section therefore concern this single retained RSF: in particular, each single-feature and full-signature model compared below is an RSF, so that the differences observed reflect the feature set alone and not a change of estimator.

Comparison with univariate features

To assess the added value of the multivariate signature, the dynamic AUC and Brier score curves of the full consensus model were compared to those obtained when each of the nine consensus features was used in isolation (Figures 3.21 and 3.22).

On the dynamic AUC (Figure 3.21), the consensus model (mean AUC = 0.759) lay above every univariate feature curve for the entire middle portion of the time window (approximately months 7 to 16), with a peak near 0.83 around months 9–10. At very early times (months 3–6), three univariate features outperformed the consensus model:

- `t1_original_glszm_GrayLevelNonUniformity_ET` (≈ 0.82 at month 3),
- `t1_original_ngtdm_Strength_NCR` (≈ 0.75 at month 3),
- `flair_wavelet-LHH_glrlm_HighGrayLevelRunEmphasis_ET` (≈ 0.66 at month 3).

Age, taken alone, had an AUC trajectory that hovered around 0.50–0.58 throughout the entire time window, briefly dropping below 0.50 in the late portion: despite being the only feature selected by all three methods and at 24/24 combinations, its univariate discrimination was

among the lowest of the nine. `t2_wavelet-LHL_firstorder_Kurtosis_ET` was the worst univariate performer, dropping below 0.50 from approximately month 8 onwards and finishing around 0.45. Several univariate features crossed the 0.50 line at various time points; the consensus model never dropped below 0.60 within the observation window. All univariate curves and the consensus curve converged to the 0.55–0.65 band at the end of follow-up.

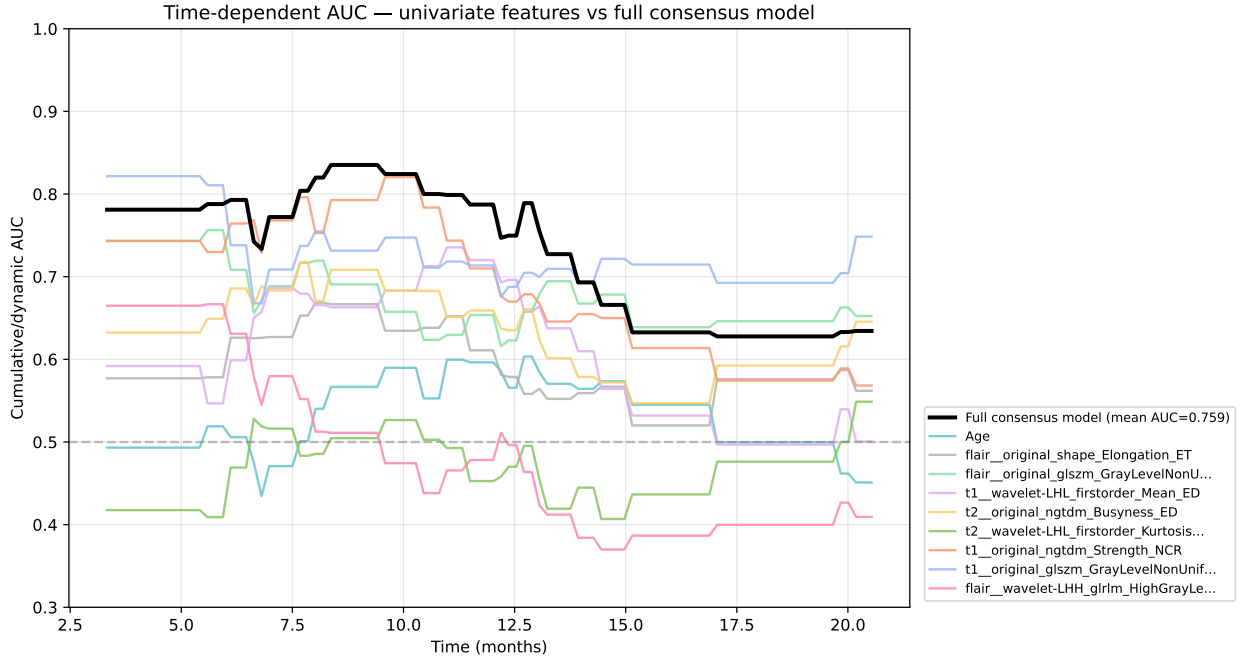


Figure 3.21: Time-dependent Cumulative/Dynamic AUC of an RSF trained on each consensus feature in isolation (coloured curves) and on the full nine-feature consensus set (black curve).

On the Brier score (Figure 3.22), the consensus model ($IBS = 0.193$) sat at or below every univariate curve for the entire time range, except briefly around month 5 where `t1_original_ngtdm_Strength_NCR` and one or two other features appeared marginally lower. The gap widened after month 7 and the consensus was typically 0.02–0.06 below the worst univariates between months 10 and 16. `flair_wavelet-LHH_glrml_HighGrayLevelRun Emphasis_ET` was the worst univariate across the middle and late portions of the window (≈ 0.28 near month 9), followed by `t2_wavelet-LHL_firstorder_Kurtosis_ET` (≈ 0.27 around month 10). `t1_original_ngtdm_Strength_NCR` was the best individual univariate predictor, sitting closest to the consensus curve throughout. The temporal pattern (low at month 5, sharp rise around months 6–8, plateau, decrease after month 15) is shared between all curves.

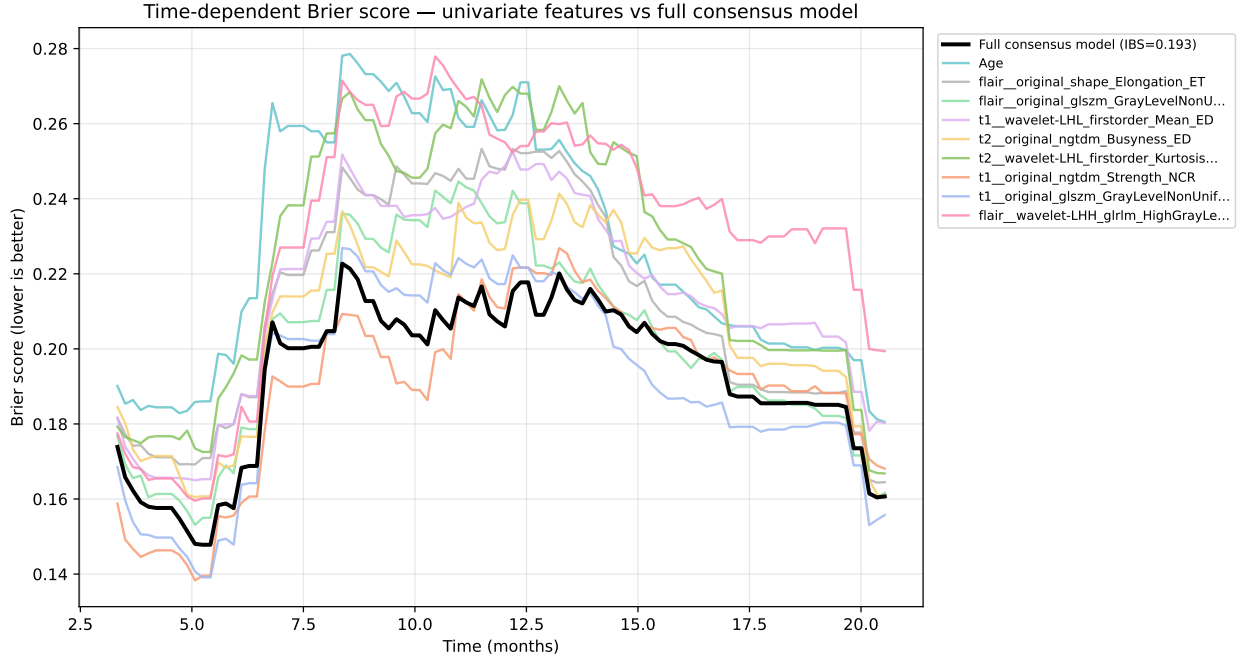


Figure 3.22: Time-dependent Brier score of an RSF trained on each consensus feature in isolation (coloured curves) and on the full nine-feature consensus set (black curve).

Permutation importance within the retained RSF

The contribution of each consensus feature to the retained RSF was quantified by permutation importance on the test set, defined as the mean drop in C-index over five permutations of the feature's values (Figure 3.23). Relative to the model's baseline C-index of 0.671, the importance was concentrated in a handful of features. `t2_original_ngtdm_Busyness_ED` dominated: permuting it lowered the C-index by $\Delta = 0.089$ (to ≈ 0.582), roughly $2.5\times$ the second-ranked feature. Three further features carried positive importance, namely `t1_original_ngtdm_Strength_NCR` ($\Delta = 0.035$), `flair_original_glszm_GrayLevelNon Uniformity_NCR` ($\Delta = 0.024$) and `t1_original_glszm_GrayLevelNonUniformity_ET` ($\Delta = 0.011$), so that almost all of the measurable importance was carried by four of the nine features.

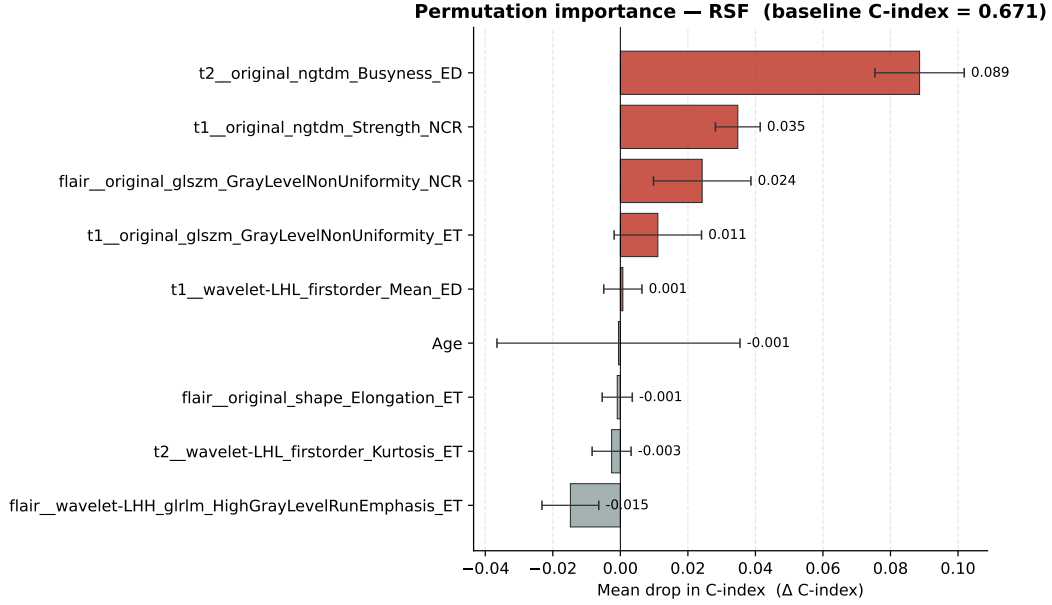


Figure 3.23: Permutation importance of the nine consensus features for the retained RSF, evaluated on the test set. The score is the mean drop in C-index over five permutations of feature values.

The remaining five features contributed little or nothing. `t1_wavelet-LHL_firstorder_Mean_ED` and `flair__original_shape_Elongation_ET` sat at $\Delta \approx 0.001$, while `t2_wavelet-LHL_firstorder_Kurtosis_ET` ($\Delta \approx -0.003$) and `flair__wavelet-LHH_glrIm_HighGrayLevelRunEmphasis_ET` ($\Delta \approx -0.015$) were slightly negative, the latter indicating that the test-set C-index improved marginally when the feature was permuted. `Age` likewise sat at $\Delta \approx -0.001$ and carried by far the widest confidence interval of the nine features (spanning from roughly -0.04 to $+0.04$). This was at odds with its standing at the selection stage, where `age` ranked first under both univariate Cox and permutation importance, was the only feature confirmed by Boruta, and was retained by all 24 method $\times$ model combinations: within the fitted nine-feature RSF, its marginal prognostic signal did not translate into a measurable contribution to discrimination. The mechanisms behind this dissociation are taken up in the discussion (Chapter 4).

Chapter 4

Discussion

4.1 Each metric answers a different clinical question

The four metrics retained in this work do not measure the same property of a model, and each of them answers a different question that a clinician would, in practice, ask of a prognostic tool. Reading the results of Section 3.2 through this clinical lens, rather than as a single performance ranking, is essential before drawing any conclusion about the suitability of these models in the management of GBM patients.

Concordance Index The C-index remains the most widely reported metric in the survival-radiomics literature, which is the first reason to keep it at the centre of this work: it places our results on a directly comparable footing with the prior GBM studies reviewed in Section 1.6. As a global scalar that requires only a relative risk score, it also has the practical advantage of being defined for every estimator in our panel, including the SSVM, which allows to compare the eight models between them (Section 2.2.5). Its interpretation is, however, deliberately modest: it tells us how often the model orders two comparable patients correctly, and nothing about whether the predicted survival probabilities themselves are accurate [50]. A limitation that needs to be taken into account is that the numerical value of the C-index also depends on the convention adopted for ties and censoring across libraries [97], so the figures reported here are not strictly interchangeable with those obtained under a different implementation, and cross-study differences of a few thousandths are not necessarily meaningful. Clinically, a high C-index is enough to support one specific use, the placement of a new patient within an existing risk ordering: an oncologist could situate the patient between two known cohorts, which is informative for prognostic counselling but does not, on its own, justify any quantitative statement about expected survival times.

Cumulative/Dynamic AUC The Cumulative/Dynamic AUC refines the C-index by resolving discrimination over time, and is the natural metric for the clinically relevant question of whether the model still ranks patients correctly at a given horizon t . Its added value here is therefore essentially diagnostic: it tells us not whether the model is useful, but at which follow-up horizons it can discriminate patients correctly. As shown in Figure 3.17, discrimination peaks in the first year and degrades sharply beyond month 15, which means that the prognostic value of these models is concentrated in the window that matters most clinically given the median survival of approximately 15 months in GBM [4]. Like the C-index, the AUC(t) only assesses the ordering of predicted risks and says nothing about the accuracy of the underlying probabilities [98, 99]. It also inherits a methodological limitation that constrained the choice of evaluation grid in Section 2.2.5: the IPCW estimator becomes unstable at the tails of the follow-up distribution, where either too few events have accrued or the at-risk set has thinned out enough to inflate the inverse-censoring weights [50].

Time-dependent Brier Score The time-dependent Brier score complements the previous two metrics on a different axis. It is the only metric in our panel that evaluates the absolute predicted survival probabilities rather than their ranking: it is a proper scoring rule, minimised only when the predicted survival probabilities coincide with the true ones, and it jointly penalises poor discrimination and poor calibration [50]. As such, it provides information that neither the C-index nor the AUC(t) can supply. Two qualifications are nevertheless in order. Assel et al. (2017) [100] argued that the Brier score is prevalence-dependent and treats false positives and false negatives symmetrically, which limits its usefulness for direct clinical decision-making; and, as already noted, the IBS is undefined for the SSVM, which excludes one of our eight estimators from this axis of comparison (Section 2.2.5). Clinically, however, the Brier score maps onto the question patients most often want answered: a model with a low $BS(t)$ at a horizon t produces survival probabilities that can be communicated to the patient and their relatives, and such a number is what enables shared decision-making about the intensity of treatment, the timing of advance care planning, and the personal decisions that follow the diagnosis.

Restricted Mean Survival Time The RMST occupies a different position from the three preceding metrics: rather than evaluating the model directly, it summarises the survival experience of the two risk groups that the model predicts, expressed on a time scale of days rather than of probabilities or rankings. Two properties make it the most clinically interpretable of our four metrics. First, the $\Delta RMST$ is read directly as a difference in life expectancy over $[0, \tau]$ [101], which is directly meaningful in the GBM setting where median survival lies within this window. Second, it remains valid when the proportional-hazards assumption is violated, a non-trivial advantage given that several of the candidate models in our panel do not assume proportional hazards in the first place [101]. Its main limitation is its dependence on the choice of τ [74, 101]: shifting the horizon would re-order the models if their predicted Kaplan-Meier curves cross within the window. The choice $\tau = 24$ months adopted in Section 2.2.5 is grounded in the median survival of approximately 15 months and on the depletion of the at-risk set beyond two years.

Read jointly, the four metrics therefore describe four complementary facets of the same model: how it orders patients globally (C-index), whether that ordering holds across time (AUC(t)), how accurate the underlying probabilities are ($BS(t)$), and how the resulting risk dichotomy translates into clinically interpretable event-free time (RMST). No single metric captures all four facets at once, which is precisely the reason for combining them in this work.

4.2 RSF selected for clinical use

Reading Table 3.7 along the four axes discussed above produces a clearer picture than any single ranking would allow, and brings out one structural fact: the eight models do not separate into a continuum of progressively better estimators, but into three groups whose ordering reflects the structural assumptions reviewed in Section 1.5.3. The five linear Cox-based configurations, together with GBSM componentwise, cluster between C-index 0.625 and 0.631, with mean AUC(t) values between 0.695 and 0.704 and IBS values between 0.198 and 0.200. The SSVM with a Gaussian kernel sits in an intermediate position at C-index 0.640 and mean AUC 0.712. The two tree-based ensembles, RSF and GBSM, sit clearly above the rest, with C-index of 0.671 and 0.662, mean AUCs of 0.739 and 0.748, and IBSs of

0.193 and 0.192. This three-tier structure is the most informative finding of the comparison, and is consistent across all three model-level metrics.

The gap between the tree-based ensembles and the linear Cox cluster can be interpreted, although the wide and mutually overlapping bootstrap intervals mean it should be read as a tendency rather than a statistically resolved separation. It indicates that the consensus features retained in Section 3.2.2 carry information that a strictly linear risk score cannot fully exploit, and that relaxing this linearity yields a measurable, although modest, gain in both discrimination and calibration. The retained GBSM configuration is informative on the source of this gain: with `max_depth = 1` (Table 3.5), its weak learners are univariate stumps, so the additive score $F(X)$ is non-linear but free of any interaction term, and yet GBSM sits essentially level with RSF across all three metrics. The two GBSM variants bracket the effect cleanly: the componentwise variant, whose learners are univariate *linear* regressions, falls back to the linear cluster, whereas the depth-one tree variant, whose learners are univariate non-linear stumps, rises to the top tier. What lifts the tree-based ensembles above the linear models is therefore primarily the modelling of non-linear univariate effects, not the capture of interactions between features: the additional step from the additive GBSM to the interaction-capable RSF (`max_depth = 4`) is worth only about 0.01 in C-index, well within the bootstrap noise.

Within the leading pair, the choice between RSF and GBSM is less straightforward and depends on which metric is given priority. On the two model-level metrics that evaluate prediction directly, GBSM has a marginal edge: its mean AUC(t) is slightly higher (0.748 versus 0.739) and its IBS is slightly lower (0.192 versus 0.193). On the C-index, however, the ordering reverses (0.671 for RSF versus 0.662 for GBSM), and the difference on the risk-stratification axis is substantial. The Kaplan-Meier curves of the two predicted risk groups (Figure 3.19) separate clearly for RSF and reach conventional significance (log-rank $p = 1.43 \times 10^{-2}$, $\Delta\text{RMST} = 192$ days), whereas GBSM fails to dichotomise patients into two distinct survival profiles: its two curves overlap throughout follow-up and even cross between 600-1000 days, the log-rank test is not significant ($p = 1.38 \times 10^{-1}$), and the ΔRMST drops to 116 days, the lowest of the entire panel. GBSM is therefore the model that predicts survival probabilities most accurately on average, but the model that translates these probabilities the least well into a clinically usable binary stratification.

This contrast captures, on our cohort, the difference between a metric-optimal model and a clinically operative one. A model intended to inform shared decision-making in GBM has to support two distinct uses: producing reliable survival probabilities for individual patients, and identifying a high-risk subgroup that could, in principle, be the target of more intensive follow-up or treatment intensification. RSF performs both tasks acceptably, with a competitive IBS and a clinically meaningful 192-day gap in restricted mean survival time between its predicted risk groups. GBSM performs only the first. In the absence of a metric that strictly dominates, the model that supports both uses is the more defensible choice, and on this basis the RSF emerges as the best overall model of the panel.

These conclusions should nevertheless be qualified by the size of the test set and the corresponding width of the bootstrap confidence intervals. Each C-index CI has a width of approximately 0.16, and every pair of CIs overlaps substantially (Section 3.2.2), which means that the ordering observed on point estimates does not reach formal statistical separation between most models. The three-tier structure described above groups several models per tier, but the fine ordering within the top tier, RSF versus GBSM, rests on a difference of

less than 0.01 on each metric. The argument for selecting RSF should therefore be read as resting on the risk-stratification axis, where the gap is large, rather than on its marginal lead in C-index, where it is not.

4.3 Comparison with the literature

The performance of the retained RSF (test-set: C-index 0.671, mean AUC(t) 0.739, IBS 0.193) is best read against the studies reviewed in Section 1.6 that, like ours, rely on conventional MRI only. The closest comparator is Duman *et al.* (2024) [85], who, on the same BraTS2020 cohort augmented to $n = 289$, reached a C-index of 0.69 and an integrated AUC at 11 months of 0.81 from a three-feature signature (age and two FLAIR-derived descriptors). Our RSF sits in the same range. Zhang *et al.* (2023) [83] provides a second held-out comparator: on 131 patients, they reported 0.667 for an RSF and 0.635 for a CoxPH, against our 0.671 and 0.625, a difference of roughly 0.004 and 0.010 respectively, close given the difference in cohort and pipeline. Taken together, these figures suggest that the upper end of what conventional BraTS2020-style MRI yields with hand-crafted features lies in a 0.67–0.69 band, and that the tree-ahead-of-Cox ordering we observe is a structural property rather than a peculiarity of our cohort.

Several studies report higher C-index, but on substrates or protocols unavailable to a BraTS2020-only pipeline. Park *et al.* (2020) [80] reached 0.70 on external validation, but by adding diffusion- and perfusion-weighted sequences (ADC, CBV) to T1ce and FLAIR and by combining the radiomic signature with age, KPS and extent of resection; the gap of approximately 0.03 to our RSF therefore reflects both the advanced sequences and the clinical variables that a BraTS2020-only pipeline does not have, and is not best read as a methodological shortfall. Hajianfar *et al.* (2023) [84] reached 0.77 on 119 patients, but as the maximum over a grid of 504 preprocessing, selection and model combinations; this is not commensurable with our single-pipeline figure, and treating the gap of approximately 0.1 as like-for-like would overstate it. The deep-learning results of Lao *et al.* (2017) [81] and the DeepSurv variant of Zhang *et al.* (2023) [83], at 0.710 and 0.732, illustrate the familiar trade-off: deep features extract additional signal but require larger samples to generalise and forfeit the biological interpretability that motivated, here, the choice of an IBSI-compliant PyRadiomics signature.

The linear baselines are more straightforward to position: our CoxPH (0.625) agrees closely with Zhang *et al.*’s CoxPH (0.635) [83] and sits well below the tree-based ensembles reported on the same cohorts. The persistence of this Cox–tree gap across studies suggests that MRI-derived radiomic features carry prognostic information that a strictly linear risk score cannot fully capture, and that a flexible estimator is among the more reliable levers for improving discrimination on a fixed feature set, a point also made by Zhang *et al.* [83] and Hajianfar *et al.* [84]. As argued in Section 4.2, our own panel localises this gain to non-linear univariate effects rather than to interactions between features or to a relaxation of the proportional-hazards form, since the additive, proportional-hazards GBSM reaches the same tier as the RSF.

4.4 The tumour core carries the prognosis

The nine features carried forward by the consensus pipeline cluster into a single clinical covariate and four radiomic families:

- a single clinical covariate (age)
- two grey-level zone descriptors (GLSZM)
- two neighbourhood intensity descriptors (NGTDM)
- three first-order or run-length descriptors derived from wavelet sub-bands
- one shape descriptor

Rather than discussing each feature in isolation, the biological narrative below groups them by family, since features within the same family share a common mathematical construction and therefore a common physical meaning when projected back onto the tumour.

Age Age is the only covariate retained by all three selection methods, the only one kept by Boruta, and it tops both the univariate Cox and the permutation rankings; on every aggregate criterion it is the dominant variable. That dominance largely dissolves inside the final model, however, where its permutation importance falls to roughly -0.001 and where it is the one feature to violate the proportional-hazards assumption ($p = 0.041$) (Table 3.4). The discrepancy is itself informative: age acts here as an unadjusted prognostic estimator rather than as a biological axis. Older patients more frequently harbour IDH-wildtype disease, present with a lower Karnofsky score, carry a heavier comorbidity burden and are less likely to undergo gross total resection [102, 103]; the BraTS2020 sub-challenge restricts evaluation to GTR patients, which attenuates the surgical confounder but not the others. Consistent with this, once IDH status, MGMT methylation and frailty are fully controlled for, the independent effect of age on survival becomes statistically negligible [104]. Its prominence should therefore be read as the expected behaviour of a composite proxy in a heterogeneously treated cohort.

Grey-level zone descriptors (GLSZM) Both retained descriptors are the Gray Level Non-Uniformity of a size-zone matrix, measured in the enhancing tumour on T1 and in the necrotic core on FLAIR (correlation 0.53). The GLSZM counts connected zones of constant grey level, and its non-uniformity rises as the voxel population fragments into many zones of differing intensity, the radiological signature of a heterogeneous region [32, 105]. The same descriptor has independently marked the worst-prognosis “heterogeneous-enhancing” phenotype [105] and ranked as the most informative predictor in a high-grade glioma survival model [106]; recovering it on two different modalities and sub-regions suggests the underlying signal is robust to both.

Neighbourhood intensity descriptors (NGTDM) NGTDM is the smallest texture class extracted, yet it supplies two consensus features, Busyness in the oedema (from T2 modality) and Strength in the necrotic core (from T1 modality), and they rank first and

second by permutation importance in the final RSF. The class measures how sharply each voxel differs from its immediate neighbourhood: Busyness rises with rapid, frequent intensity changes, describing a restless, fine-grained texture, whereas Strength rises with the presence of a few large, well-differentiated intensity primitives [32]. Both therefore index the abrupt local transitions of a disorganised microstructure, the same heterogeneity that tracks aggressiveness in glioblastoma [27]. External support exists at the class level: Hajianfar et al. retained NGTDM Coarseness, the canonical descriptor of this matrix, among the texture features most consistently selected across seven feature-selection methods in a glioblastoma survival model, with a busier, less coarse enhancing-tumour texture associated with shorter survival [84]. We are aware of no study isolating Busyness or Strength specifically, however, so the interpretation of these two individual metrics remains tentative.

Wavelet-derived descriptors (first-order and run-length) Three features share the same construction, extraction from a directional wavelet sub-band rather than the native image, and are best read together: the Mean of the oedema (T1, LHL), the Kurtosis of the enhancing tumour (T2, LHL), and a High Gray Level Run Emphasis in the enhancing tumour (FLAIR, LHH). The two first-order statistics summarise the filtered intensity distribution without any spatial information [27, 32]; we have no glioblastoma-specific prognostic reference for either, so their biological reading is offered only as a plausible interpretation: the Mean reflects the average signal of the mixed vasogenic-and-infiltrative oedema [6, 28], while Kurtosis captures the peakedness of the histogram and is sensitive to a minority of extreme voxels, possibly focal necrosis or haemorrhage within the rim. High Gray Level Run Emphasis, by contrast, increases where bright voxels form spatially extended runs and has been reported as prognostic in diffuse glioma, stratifying overall survival in both a training and a validation cohort, although in lower-grade rather than grade-4 tumours [107]. This run-length feature is also the most weakly selected of the entire set, retained by only two of the twenty-four combinations, so it should be regarded as the least robust descriptor among the nine.

Shape The only shape descriptor is the Elongation of the enhancing tumour, computed from the mask alone and so independent of intensity and modality. Its name is misleading: in PyRadiomics, Elongation is the ratio of the minor to the major principal axis, bounded in $[0, 1]$, so a value near 1 denotes a rounded, near-spherical shape and a value near 0 an elongated one [32]. A high value thus marks a more regular tumour. Shape regularity is among the most reproducibly prognostic morphological properties in glioblastoma, although the published evidence rests mainly on the closely related sphericity descriptor: Pérez-Beteta et al. established surface regularity as an independent predictor of survival and of the benefit from gross total resection [108], a result since replicated with more spherical tumours associated with longer survival [109]. Both sphericity and elongation also recur among the shape features most frequently selected in independent glioblastoma survival models [83].

An anatomical reading Across families, the prognostic signal concentrates in the tumour core: four of the eight radiomic features fall in the enhancing tumour and two more in the necrotic core, consistent with prior glioblastoma radiomics in which the core carries most of the prognostic information [78]. Two absences sharpen the picture, but both are better read as properties of the pipeline than as biological findings. T1ce contributes no feature at all despite defining the enhancing region; rather than reflecting an absence of prognostic texture,

this is the expected consequence of the alphabetical tie-break used in the redundancy filter (Section 2.2.3), which retains, within each correlated pair, the feature whose name comes first. Since `t1ce` sorts behind both `flair` and `t1` (ahead only of `t2`), any `T1ce` feature correlated above the 0.75 threshold with a `FLAIR` or `T1` feature is dropped in its favour; given how strongly `T1` and `T1ce` textures co-vary, this rule alone is sufficient to explain the complete absence of `T1ce`, which is therefore most plausibly an artefact of the tie-break rather than evidence that its enhancement texture is uninformative. `GLCM` and `GLDM`, the two largest classes, are likewise absent, but for different reasons: `GLDM` is heavily redundant with `GLRLM` and is largely removed at the correlation filter (Figure 3.3), whereas `GLCM` survives that step yet rarely passes supervised selection, an absence better read as “not prognostic in this cohort” than as redundancy. All three readings should be confirmed against the filter log and the per-modality co-occurrence of correlated pairs (Figure 3.2) rather than asserted.

4.5 Limitations

The results presented in this work rest on a single cohort, and their generalisation beyond it remains to be established. All models were trained and evaluated on BraTS2020 through an internal 80/20 split (Section 2.2.3); no independent external cohort was used to confirm that the retained RSF and its nine-feature signature transfer to images acquired under different scanners, field strengths and protocols. This is the most important limitation of the study, because radiomic features are known to be sensitive to acquisition conditions (Section 2.2.1), and a signature calibrated on one multi-center collection may still fail to reproduce on another. Two further properties of the cohort compound this concern. First, the test set comprises only 47 patients and the bootstrap confidence intervals are wide (approximately 0.16 on the C-index) and overlap across the panel; as discussed in Section 4.2, the ordering of the models does not reach formal statistical separation, and the selection of the RSF rests on the risk-stratification axis rather than on its marginal lead in C-index. Second, the censoring rate is almost null (0.4%, a single patient in the training set), because overall survival in BraTS2020 is provided almost only for cases in which the event was observed; the survival framework is therefore formally applicable but barely mobilised, and the capacity of the models to handle censored observations, the very property that justifies survival analysis over a direct regression on event time, is essentially untested here.

The prognostic model is also incomplete with respect to the clinical and molecular information routinely used in neuro-oncology. Age was the only non-imaging covariate retained, since resection status was missing for almost half of the cohort and no molecular annotation was available (Section 2.1). Yet `MGMT` promoter methylation and `IDH` status carry independent prognostic value in glioblastoma, and the extent of resection and the Karnofsky performance status are well-established predictors (Section 1.6). The radiomic signature is thus asked to carry prognostic signal that, in a complete clinical workup, would be shared with these variables.

Several choices in the imaging pipeline further bound the scope of the results. The analysis was restricted to the four conventional sequences available in BraTS2020 (`T1`, `T1ce`, `T2`, `FLAIR`); diffusion- and perfusion-weighted maps, which have been shown to push discrimination further (Park et al., 2020 [80]), were not available and could not contribute. Intensity standardisation relied on an intra-image z-score, and no bias-field correction (e.g. `N4ITK`) was applied beforehand (Section 2.2.1), so any residual low-frequency inhomogeneity may have

propagated into the texture features. Feature extraction was moreover performed on a single segmentation, the BraTS reference annotation, without assessing the stability of the features to plausible variations of the contour; since radiomic descriptors, and texture descriptors in particular, are known to be sensitive to segmentation, the robustness of the nine-feature signature to inter-observer or automated re-segmentation was not characterised. Finally, the deliberate restriction to hand-crafted features (Section 1.3), justified by interpretability, also sets a ceiling on the achievable performance relative to learned representations.

The feature-selection procedure itself carries methodological caveats that temper the interpretation of the final signature. Within each correlated pair, the redundancy filter retained the feature first in alphabetical order (Section 2.2.3); this rule has no statistical justification and introduces a systematic bias in favour of the modalities and classes whose names come earlier in the PyRadiomics convention (FLAIR before T1, T1ce and T2; first-order and GLCM before later classes). The consensus criterion was also relatively permissive: a feature qualified by being selected in at least two of the twenty-four configurations, a threshold that a single method applied to two models already satisfies, so the criterion is best understood as a filter against single-configuration artefacts rather than as a strong multi-method consensus; indeed, seven of the nine retained features were selected by a single method. Lastly, the entire selection was carried out once, on a single train/test partition, and the stability of the signature under resampling was not assessed, so it is not known how much the nine features would change had the split been drawn differently.

A further methodological caveat concerns the stage at which the final model was chosen and the way its performance was reported. The eight candidates were carried into the final evaluation with their hyperparameters fixed on the training set alone (Section 2.2.4), so no information leaked during calibration; the comparison that singled out the RSF, however, was itself carried out on the held-out test set. All four axes were computed on the test set for every candidate, the three metrics and, decisively, the risk-group separation on which the choice primarily rested (Section 4.2), and the model that best dichotomised the two risk groups on that particular set ($p = 1.43 \times 10^{-2}$, $\Delta\text{RMST} = 192$ days) was retained, after which the same test-set figures were reported as its final performance. Using the test set both to select the winning model and to quantify it renders these figures optimistically biased: when several candidates of comparable expected performance are ranked on a single finite test set, the apparent superiority of the top-ranked one partly reflects favourable sampling fluctuation rather than a genuine advantage, a selection effect that inflates the reported result. The concern is sharpened by the closeness of the panel: the eight test-set C-indices span only ≈ 0.046 within mutually overlapping bootstrap intervals, and all seven significant log-rank p -values lie in the narrow 0.01–0.05 band, so the models are not statistically distinguishable on this cohort and the choice of the RSF rests on differences of the order of the sampling noise. A more rigorous protocol would have conducted the entire comparison, including the risk-group separation, here assessed only on the test set, within the cross-validation on the training data, locked the single best model on that basis, and only then evaluated that one model, once, on the test set, so that the reported test-set performance would constitute a genuinely held-out and therefore unbiased estimate. The true generalisation performance of the retained model, and in particular the robustness of its risk dichotomy, is consequently likely to be somewhat weaker than the headline test-set values suggest.

Finally, the biological and anatomical interpretation of the retained signature (Section 4.4) was carried out on the basis of the radiomics and neuro-oncology literature alone, without direct input from a clinician. The proposed readings, the concentration of the prognostic

signal in the tumour core, and the tentative biological meaning attached to individual texture descriptors, would have been strengthened by discussion with specialists who interpret these images and treat these patients in practice: principally a neuroradiologist and a neuro-oncologist or radiation oncologist, with a neurosurgeon contributing an anatomical and operative perspective. Such expertise would help to assess the clinical plausibility of the associations, to separate genuine biological signal from the pipeline-induced artefacts that the present work could only flag rather than resolve (Section 4.4), and to guard against over-interpreting features whose statistical selection does not, in itself, establish any mechanistic link to tumour biology.

4.6 Future perspectives

The most immediate continuation of this work is external validation. Several large public glioblastoma cohorts that were not exploited here are available and would allow the retained RSF and its nine-feature signature to be tested on genuinely independent data: the UPenn-GBM collection [110], the currently largest public glioblastoma dataset with roughly 630 patients and accompanying clinical, molecular and advanced-imaging information; the UCSF-PDGM dataset [111]; and the smaller RHUH-GBM cohort [112]. Validating the locked signature on one or more of these collections would directly address the single-cohort limitation, quantify the loss of discrimination across scanners and protocols, and, because these cohorts include genuinely censored follow-up, also exercise the survival machinery that the near-complete BraTS2020 data left idle.

A second direction is to enrich the model along the two axes that the present cohort constrained. On the clinical side, repeating the analysis on a cohort with molecular and surgical annotation (MGMT, IDH, extent of resection, KPS) would make it possible to measure the incremental prognostic value of the radiomic signature over a complete clinical-molecular baseline, which is ultimately the quantity that determines its clinical usefulness. On the imaging side, an explicit multi-centre harmonisation step would improve transferability: harmonising the radiomic features with ComBat or one of its derivatives (M-ComBat, AutoComBat) [113], and adding an N4 bias-field correction [88] upstream of normalisation, would mitigate the centre effects that a single-cohort pipeline does not control and that an external validation is likely to expose.

Finally, the radiomics pipeline itself could be made more robust. The alphabetical tie-break in the redundancy filter should be replaced by a principled rule, retaining within each correlated pair the feature with the highest variance, the lowest mean correlation with the remaining set, or the strongest association with the target; and the consensus criterion should be tightened once a larger cohort makes a stricter threshold affordable. Embedding the whole selection within nested resampling, and reporting the bootstrap stability of the signature together with the reproducibility of each feature under segmentation perturbation (for instance using automated segmentations as an alternative to the reference contour), would turn the nine-feature set from a single-split result into a quantifiably stable signature. Pursued together, external validation, clinical-molecular enrichment, harmonisation and a reproducibility-aware selection would bring an interpretable, conventional-MRI prognostic model substantially closer to the standard required for clinical translation.

Conclusion

Glioblastoma is managed patient by patient, yet that individualisation still rests largely on chronological age, while the quantitative content of the MRI examinations acquired throughout care goes mostly unread. This thesis set out to ask whether that routinely available imaging could be turned, automatically and reproducibly, into an interpretable estimate of a patient’s prognosis. Working from the four conventional sequences (T1, T1ce, T2 and FLAIR) of the BraTS2020 cohort alone, it extracted 10,203 hand-crafted radiomic features per patient across three tumour sub-regions, reduced them through a multi-method consensus procedure to a compact signature of nine variables, the patient’s age and eight radiomic descriptors, and compared eight survival models under a single protocol before retaining, on a strictly held-out test set, the one most suited to clinical use.

That model, a Random Survival Forest, reached a concordance index of 0.671, a mean time-dependent AUC of 0.739 and an integrated Brier score of 0.193 on the test set, and it separated the cohort into two significantly distinct prognostic groups, a 192-day difference in restricted mean survival time, log-rank $p = 1.43 \times 10^{-2}$. These figures place a model built on conventional MRI alone on a par with the conventional-MRI state of the art, and they do so without sacrificing transparency, since every variable retained carries a direct and traceable biological meaning. Two messages follow. The first is that interpretability and competitive performance need not be traded against one another in glioblastoma prognostication: a parsimonious, fully legible signature, drawn from imaging that virtually every center already acquires, is enough to stratify patients at a clinically meaningful level. The second concerns how such a model should be chosen, no estimator dominated the panel on every metric, and the model was retained not for topping a single number but because it answered two distinct clinical questions at once, producing reliable individual survival probabilities while isolating a high-risk subgroup, where the alternatives answered only one. Reading performance through the lens of the decisions a clinician actually faces, rather than as a single ranking, is itself part of what the work contributes.

The second objective, identifying interpretable imaging biomarkers, converged on a coherent picture. The prognostic signal concentrated in the tumour core: most of the retained radiomic descriptors fell in the enhancing tumour and the necrotic core, and they captured intratumoural heterogeneity, the fragmented, abruptly varying texture of a disorganised microstructure that tracks tumour aggressiveness, together with the regularity of the tumour’s shape, itself among the most reproducibly prognostic morphological properties in glioblastoma. These readings, however, rest on the radiomics and neuro-oncology literature alone, with no direct clinical input, and so remain tentative; confirming them with the specialists who interpret these images and treat these patients, principally a neuroradiologist and a neuro-oncologist, would be a natural way to take the analysis further.

Taken together, these results offer a partial but concrete answer to the question raised at the outset. An automated, quantitative and interpretable prognostic estimate can be derived from a routine MRI examination, and the risk stratification it affords bears directly on the clinical uses that motivated the work: calibrating the intensity of imaging follow-up, enriching the

high-risk arms of clinical trials, and supporting a quantified, and, where the patient wishes it, candid, discussion of prognosis. The principal reservation is equally clear, and points to the natural next step: the signature was established and tested within a single cohort, and its transfer across scanners, protocols and centres remains to be shown. External validation on an independent collection, ideally one enriched with the molecular and surgical variables that BraTS2020 lacks, is what would carry an interpretable, conventional-MRI prognostic model from proof of concept towards a tool fit for the clinic. The present work shows that such a model is within reach, and that clinically meaningful prognostic information can be drawn from the very imaging glioblastoma patients already undergo.

Bibliography

- [1] Anthropic, “Claude opus 4.7.” Large language model.
- [2] DeepL SE, “DeepL translator.”
- [3] Corporation for Digital Scholarship, “Zotero.” Version 9.
- [4] C. McKinnon, M. Nandhabalan, S. A. Murray, and P. Plaha, “Glioblastoma: clinical presentation, diagnosis, and management,” *BMJ*, p. n1560, July 2021.
- [5] M. Davis, “Glioblastoma: Overview of Disease and Treatment,” *Clinical Journal of Oncology Nursing*, vol. 20, pp. S2–S8, Oct. 2016.
- [6] H.-G. Wirsching and M. Weller, “Glioblastoma,” in *Malignant Brain Tumors* (J. Moliterno Gunel, J. M. Piepmeier, and J. M. Baehring, eds.), pp. 265–288, Cham: Springer International Publishing, 2017.
- [7] I. A. Tewarie, J. T. Senders, S. Kremer, S. Devi, W. B. Gormley, O. Arnaout, T. R. Smith, and M. L. D. Broekman, “Survival prediction of glioblastoma patients—are we there yet? A systematic review of prognostic modeling for glioblastoma and its clinical potential,” *Neurosurgical Review*, vol. 44, pp. 2047–2057, Aug. 2021.
- [8] H. Krenzlin, D. Jankovic, C. Alberter, D. Kalasauskas, C. Westphalen, F. Ringel, and N. Keric, “Frailty in Glioblastoma Is Independent From Chronological Age,” *Frontiers in Neurology*, vol. 12, 2021.
- [9] C. McCague, S. Ramlee, M. Reinius, I. Selby, D. Hulse, P. Piyatissa, V. Bura, M. Crispin-Ortuzar, E. Sala, and R. Woitek, “Introduction to radiomics for a clinical audience,” *Clinical Radiology*, vol. 78, pp. 83–98, Feb. 2023.
- [10] E. Avery, P. C. Sanelli, M. Aboian, and S. Payabvash, “Radiomics: A Primer on Processing Workflow and Analysis,” *Seminars in Ultrasound, CT and MRI*, vol. 43, pp. 142–146, Apr. 2022.
- [11] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. S. Kirby, J. B. Freymann, K. Farahani, and C. Davatzikos, “Advancing The Cancer Genome Atlas glioma MRI collections with expert segmentation labels and radiomic features,” *Scientific Data*, vol. 4, p. 170117, Sept. 2017.
- [12] S. Bakas, M. Reyes, *et al.*, “Identifying the Best Machine Learning Algorithms for Brain Tumor Segmentation, Progression Assessment, and Overall Survival Prediction in the BRATS Challenge,” 2018. Version Number: 3.
- [13] B. H. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, L. Lanczi, E. Gerstner, M.-A. Weber, T. Arbel, B. B. Avants, N. Ayache, P. Buendia, D. L. Collins, N. Cordier, J. J. Corso, A. Criminisi, T. Das, H. Delingette, C. Demiralp, C. R. Durst, M. Dojat, S. Doyle, J. Festa, F. Forbes, E. Geremia, B. Glocker, P. Golland, X. Guo, A. Hamamci, K. M. Iftekharuddin, R. Jena, N. M. John, E. Konukoglu, D. Lashkari, J. A. Mariz, R. Meier, S. Pereira, D. Precup,

- S. J. Price, T. R. Raviv, S. M. S. Reza, M. Ryan, D. Sarikaya, L. Schwartz, H.-C. Shin, J. Shotton, C. A. Silva, N. Sousa, N. K. Subbanna, G. Szekely, T. J. Taylor, O. M. Thomas, N. J. Tustison, G. Unal, F. Vasseur, M. Wintermark, D. H. Ye, L. Zhao, B. Zhao, D. Zikic, M. Prastawa, M. Reyes, and K. Van Leemput, “The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS),” *IEEE Transactions on Medical Imaging*, vol. 34, pp. 1993–2024, Oct. 2015.
- [14] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, “Segmentation Labels for the Pre-operative Scans of the TCGA-GBM collection,” 2017.
- [15] S. Bakas, H. Akbari, A. Sotiras, M. Bilello, M. Rozycki, J. Kirby, J. Freymann, K. Farahani, and C. Davatzikos, “Segmentation Labels for the Pre-operative Scans of the TCGA-LGG collection,” 2017.
- [16] R.-J. M. Van Geuns, P. A. Wielopolski, H. G. De Bruin, B. J. Rensing, P. M. Van Ooijen, M. Hulshoff, M. Oudkerk, and P. J. De Feyter, “Basic principles of magnetic resonance imaging,” *Progress in Cardiovascular Diseases*, vol. 42, pp. 149–156, Sept. 1999.
- [17] S. Alzola-Aldamizetxebarria, L. Fernández-Méndez, D. Padro, J. Ruíz-Cabello, and P. Ramos-Cabrera, “A Comprehensive Introduction to Magnetic Resonance Imaging Relaxometry and Contrast Agents,” *ACS Omega*, vol. 7, pp. 36905–36917, Oct. 2022.
- [18] W. A. Gibby, “Basic principles of magnetic resonance imaging,” *Neurosurgery Clinics of North America*, vol. 16, pp. 1–64, Jan. 2005.
- [19] S. Hiremath, “Magnetic Resonance Imaging (MRI): Principles, System, Engineering Considerations for Image Optimization, and Medical Applications,” *Progress in Medical Sciences*, pp. 1–4, Feb. 2023.
- [20] P. Sprawls, *Magnetic resonance imaging: principles, methods, and techniques*. Madison, Wis.: Medical Physics Pub., 2000. OCLC: 44040528.
- [21] B. Kastler and P. Anstett, *Comprendre l’IRM: manuel d’auto-apprentissage*. Imagerie médicale diagnostic, Issy-les-Moulineaux [France]: Elsevier Masson, 7ème éd. entièrement rév éd., 2011.
- [22] L. Dricot, “LGBIO2050: MRI,” 2024.
- [23] M. Gaeta, K. Galletta, M. Cavallaro, E. Mormina, M. T. Cannizzaro, L. R. M. Lanzafame, T. D’Angelo, A. Blandino, S. L. Vinci, and F. Granata, “T1 relaxation: Chemo-physical fundamentals of magnetic resonance imaging and clinical applications,” *Insights into Imaging*, vol. 15, p. 200, Aug. 2024.
- [24] V. D. Köchli, B. Marinček, and D. Weishaupt, *How does MRI work? an introduction to the physics and function of magnetic resonance imaging*. Berlin [u.a.]: Springer, 2. ed., corr. 2. print ed., 2008.
- [25] B. Bonato, L. Nanni, and A. Bertoldo, “Advancing Precision: A Comprehensive Review of MRI Segmentation Datasets from BraTS Challenges (2012–2025),” *Sensors*, vol. 25, p. 1838, Mar. 2025.

- [26] A. C. Tan, D. M. Ashley, G. Y. López, M. Malinzak, H. S. Friedman, and M. Khasraw, “Management of glioblastoma: State of the art and future directions,” *CA: A Cancer Journal for Clinicians*, vol. 70, pp. 299–312, July 2020.
- [27] U. Schick, F. Lucia, G. Dissaux, D. Visvikis, B. Badic, I. Masson, O. Pradier, V. Bourbonne, and M. Hatt, “MRI-derived radiomics: methodology and clinical applications in the field of pelvic oncology,” *The British Journal of Radiology*, vol. 92, p. 20190105, Dec. 2019.
- [28] A. Hawkins-Daarud, R. C. Rockne, A. R. A. Anderson, and K. R. Swanson, “Modeling Tumor-Associated Edema in Gliomas during Anti-Angiogenic Therapy and Its Impact on Imageable Tumor,” *Frontiers in Oncology*, vol. 3, 2013.
- [29] J. E. Van Timmeren, D. Cester, S. Tanadini-Lang, H. Alkadhi, and B. Baessler, “Radiomics in medical imaging—“how-to” guide and critical reflection,” *Insights into Imaging*, vol. 11, p. 91, Dec. 2020.
- [30] W. Zhang, Y. Guo, and Q. Jin, “Radiomics and Its Feature Selection: A Review,” *Symmetry*, vol. 15, p. 1834, Sept. 2023.
- [31] P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R. G. Van Stiphout, P. Granton, C. M. Zegers, R. Gillies, R. Boellard, A. Dekker, and H. J. Aerts, “Radiomics: Extracting more information from medical images using advanced feature analysis,” *European Journal of Cancer*, vol. 48, pp. 441–446, Mar. 2012.
- [32] PyRadiomics Community, “Pyradiomics documentation.” <https://pyradiomics.readthedocs.io/en/latest/>, 2026. Accessed: 2026-05-20.
- [33] M. E. Mayerhoefer, A. Materka, G. Langs, I. Häggström, P. Szczypiński, P. Gibbs, and G. Cook, “Introduction to Radiomics,” *Journal of Nuclear Medicine*, vol. 61, pp. 488–495, Apr. 2020.
- [34] S. Majumder, S. Katz, D. Kontos, and L. Roshkovan, “State of the art: radiomics and radiomics-related artificial intelligence on the road to clinical translation,” *BJR/Open*, vol. 6, p. tzad004, Dec. 2023.
__eprint: <https://academic.oup.com/bjro/article-pdf/6/1/tzad004/58903343/tzad004.pdf>.
- [35] J. E. Park, S. Y. Park, H. J. Kim, and H. S. Kim, “Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives,” *Korean Journal of Radiology*, vol. 20, no. 7, p. 1124, 2019.
- [36] L. Rundo and C. Militello, “Image biomarkers and explainable AI: handcrafted features versus deep learned features,” *European Radiology Experimental*, vol. 8, p. 130, Nov. 2024.
- [37] V. Berisha, C. Krantsevich, P. R. Hahn, S. Hahn, G. Dasarathy, P. Turaga, and J. Liss, “Digital medicine and the curse of dimensionality,” *npj Digital Medicine*, vol. 4, p. 153, Oct. 2021.
- [38] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, Mar. 2003.

- [39] A. Bommert, T. Welchowski, M. Schmid, and J. Rahnenführer, “Benchmark of filter methods for feature selection in high-dimensional gene expression survival data,” *Briefings in Bioinformatics*, vol. 23, p. bbab354, Jan. 2022.
- [40] C. Chen, Y. Tsai, F. Chang, and W. Lin, “Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results,” *Expert Systems*, vol. 37, p. e12553, Oct. 2020.
- [41] M. Lu, “Embedded feature selection accounting for unknown data heterogeneity,” *Expert Systems with Applications*, vol. 119, pp. 350–361, Apr. 2019.
- [42] E. I. Styliara, L. G. Astrakas, G. Alexiou, V. G. Xydis, A. Zikou, G. Kafritsas, S. Voulgaris, and M. I. Argyropoulou, “Survival Outcome Prediction in Glioblastoma: Insights from MRI Radiomics,” *Current Oncology*, vol. 31, pp. 2233–2243, Apr. 2024.
- [43] scikit-learn developers, “Permutation feature importance — scikit-learn documentation.” https://scikit-learn.org/stable/modules/permutation_importance.html, 2026. Accessed: 2026-05-20.
- [44] M. B. Kursa and W. R. Rudnicki, “Feature Selection with the Boruta Package,” *Journal of Statistical Software*, vol. 36, no. 11, pp. 1–13, 2010.
- [45] A. D’Agostino, “Feature selection with boruta in python.” <https://medium.com/data-science/feature-selection-with-boruta-in-python-676e3877e596>, Nov. 2021. Accessed: 2026-05-20.
- [46] A. Kumar, V. Gunjan, Y. Hu, and S. Senatore, *Proceedings of the 4th International Conference on Data Science, Machine Learning and Applications: ICDSMLA 2022, 26–27 December, Hyderabad, India*. Lecture Notes in Electrical Engineering, Springer Nature Singapore, 2023.
- [47] A. Indrayan and C. B. Tripathi, “Survival Analysis: Where, Why, What and How?,” *Indian Pediatrics*, vol. 59, pp. 74–79, Jan. 2022.
- [48] T. G. Clark, M. J. Bradburn, S. B. Love, and D. G. Altman, “Survival Analysis Part I: Basic concepts and first analyses,” *British Journal of Cancer*, vol. 89, pp. 232–238, July 2003.
- [49] R. Flynn, “Survival analysis,” *Journal of Clinical Nursing*, vol. 21, pp. 2789–2797, Oct. 2012.
- [50] scikit-survival developers, “scikit-survival user guide.” https://scikit-survival.readthedocs.io/en/stable/user_guide/index.html, 2026. Accessed: 2026-05-20.
- [51] A. Hazra and N. Gogtay, “Biostatistics series module 9: Survival analysis,” *Indian Journal of Dermatology*, vol. 62, no. 3, p. 251, 2017.
- [52] D. Piovani, G. K. Nikolopoulos, and S. Bonovas, “Pitfalls and perils of survival analysis under incorrect assumptions: the case of COVID-19 data,” *Biomédica*, vol. 41, pp. 21–28, Oct. 2021.
- [53] F. Bella, F. Gulisano, and V. Gagliardi, “Safety analysis of motorcyclists’ overtaking maneuvers,” *Accident Analysis & Prevention*, vol. 208, p. 107809, Dec. 2024.

- [54] D. Zhang, “Survival analysis for customer retention.” <https://tinyheero.github.io/2016/05/12/survival-analysis.html>, 2016. Accessed: 2026-05-20.
- [55] Wikipedia contributors, “Kaplan–meier estimator.” https://en.wikipedia.org/wiki/Kaplan%E2%80%93Meier_estimator, 2026. Accessed: 2026-05-20.
- [56] J. T. Rich, J. G. Neely, R. C. Paniello, C. C. J. Voelker, B. Nussenbaum, and E. W. Wang, “A practical guide to understanding Kaplan-Meier curves,” *Otolaryngology–Head and Neck Surgery*, vol. 143, pp. 331–336, Sept. 2010.
- [57] Wikipedia contributors, “Nelson–aalen estimator.” https://en.wikipedia.org/wiki/Nelson%E2%80%93Aalen_estimator, 2026. Accessed: 2026-05-20.
- [58] S. Abd ElHafeez, G. D’Arrigo, D. Leonardis, M. Fusaro, G. Tripepi, and S. Roumeliotis, “Methods to Analyze Time-to-Event Data: The Cox Regression Analysis,” *Oxidative Medicine and Cellular Longevity*, vol. 2021, p. 1302811, Jan. 2021.
- [59] M. J. Bradburn, T. G. Clark, S. B. Love, and D. G. Altman, “Survival Analysis Part II: Multivariate data analysis – an introduction to concepts and methods,” *British Journal of Cancer*, vol. 89, pp. 431–436, Aug. 2003.
- [60] M. M. Nuño and D. L. Gillen, “Alternative Sampling Designs for Time-to-Event Data With Applications to Biomarker Discovery in Alzheimer’s Disease,” in *Handbook of Statistics*, vol. 36, pp. 105–166, Elsevier, 2017.
- [61] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, “Regularization Paths for Cox’s Proportional Hazards Model via Coordinate Descent,” *Journal of Statistical Software*, vol. 39, no. 5, 2011.
- [62] H. Zou and T. Hastie, “Regularization and Variable Selection Via the Elastic Net,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, pp. 301–320, Apr. 2005.
- [63] R. Tibshirani, “Regression Shrinkage and Selection Via the Lasso,” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 58, pp. 267–288, Jan. 1996.
- [64] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, “Random survival forests,” *The Annals of Applied Statistics*, vol. 2, Sept. 2008.
- [65] H. Ishwaran and M. Lu, “Random Survival Forests,” in *Wiley StatsRef: Statistics Reference Online* (R. S. Kenett, N. T. Longford, W. W. Piegorsch, and F. Ruggeri, eds.), pp. 1–13, Wiley, 1 ed., Feb. 2019.
- [66] M. Berkowitz, R. M. Altman, and T. M. Loughin, “Random forests for survival data: which methods work best and under what conditions?,” *The International Journal of Biostatistics*, vol. 20, pp. 315–345, Dec. 2024.
- [67] J. H. Friedman, “Greedy function approximation: A gradient boosting machine,” *The Annals of Statistics*, vol. 29, Oct. 2001.
- [68] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Frontiers in Neuroinformatics*, vol. 7, 2013.

- [69] APXML, “Survival analysis with gradient boosting.” <https://apxml.com/courses/mastering-gradient-boosting-algorithms/chapter-9-gradient-boosting-specialized-tasks/survival-analysis-boosting>, 2026. Accessed: 2026-05-20.
- [70] V. Van Belle, K. Pelckmans, S. Van Huffel, and J. A. Suykens, “Support vector methods for survival analysis: a comparison between ranking and regression approaches,” *Artificial Intelligence in Medicine*, vol. 53, pp. 107–118, Oct. 2011.
- [71] S. Pölsterl, N. Navab, and A. Katouzian, “Fast Training of Support Vector Machines for Survival Analysis,” in *Machine Learning and Knowledge Discovery in Databases* (A. Appice, P. P. Rodrigues, V. Santos Costa, J. Gama, A. Jorge, and C. Soares, eds.), vol. 9285, pp. 243–259, Cham: Springer International Publishing, 2015. Series Title: Lecture Notes in Computer Science.
- [72] G. Dong, L. Mao, B. Huang, M. Gamalo-Siebers, J. Wang, G. Yu, and D. C. Hoaglin, “The inverse-probability-of-censoring weighting (IPCW) adjusted win ratio statistic: an unbiased estimator in the presence of independent censoring,” *Journal of Biopharmaceutical Statistics*, vol. 30, pp. 882–899, Sept. 2020.
- [73] H. Uno, T. Cai, M. J. Pencina, R. B. D’Agostino, and L. J. Wei, “On the C-statistics for evaluating overall adequacy of risk prediction procedures with censored survival data,” *Statistics in Medicine*, vol. 30, pp. 1105–1117, May 2011.
- [74] A. Schenk, V. Basten, and M. Schmid, “Modeling the restricted mean survival time using pseudo-value random forests,” *Statistics in Medicine*, vol. 44, no. 5, p. e70031, 2025.
- [75] J. J. Z. Liao, G. F. Liu, and W.-C. Wu, “Dynamic RMST curves for survival analysis in clinical trials,” *BMC Medical Research Methodology*, vol. 20, p. 218, Aug. 2020.
- [76] P. Kickingereder, S. Burth, A. Wick, M. Götz, O. Eidel, H.-P. Schlemmer, K. H. Maier-Hein, W. Wick, M. Bendszus, A. Radbruch, and D. Bonekamp, “Radiomic Profiling of Glioblastoma: Identifying an Imaging Predictor of Patient Survival with Improved Performance over Established Clinical and Radiologic Risk Models,” *Radiology*, vol. 280, pp. 880–889, Sept. 2016.
- [77] S. Bae, Y. S. Choi, S. S. Ahn, J. H. Chang, S.-G. Kang, E. H. Kim, S. H. Kim, and S.-K. Lee, “Radiomic MRI Phenotyping of Glioblastoma: Improving Survival Prediction,” *Radiology*, vol. 289, pp. 797–806, Dec. 2018.
- [78] P. Kickingereder, U. Neuberger, D. Bonekamp, P. L. Piechotta, M. Götz, A. Wick, M. Sill, A. Kratz, R. T. Shinohara, D. T. W. Jones, A. Radbruch, J. Muschelli, A. Unterberg, J. Debus, H.-P. Schlemmer, C. Herold-Mende, S. Pfister, A. Von Deimling, W. Wick, D. Capper, K. H. Maier-Hein, and M. Bendszus, “Radiomic subtyping improves disease stratification beyond key molecular, clinical, and standard imaging characteristics in patients with glioblastoma,” *Neuro-Oncology*, vol. 20, pp. 848–857, May 2018.
- [79] U. Baid, S. U. Rane, S. Talbar, S. Gupta, M. H. Thakur, A. Moiyadi, and A. Mahajan, “Overall Survival Prediction in Glioblastoma With Radiomic Features Using Machine Learning,” *Frontiers in Computational Neuroscience*, vol. 14, 2020.

- [80] J. E. Park, H. S. Kim, Y. Jo, R.-E. Yoo, S. H. Choi, S. J. Nam, and J. H. Kim, “Radiomics prognostication model in glioblastoma using diffusion- and perfusion-weighted MRI,” *Scientific Reports*, vol. 10, p. 4250, Mar. 2020.
- [81] J. Lao, Y. Chen, Z.-C. Li, Q. Li, J. Zhang, J. Liu, and G. Zhai, “A Deep Learning-Based Radiomics Model for Prediction of Survival in Glioblastoma Multiforme,” *Scientific Reports*, vol. 7, p. 10353, Sept. 2017.
- [82] A. Chaddad, M. Zhang, C. Desrosiers, and T. Niazi, “Deep Radiomic Features from MRI Scans Predict Survival Outcome of Recurrent Glioblastoma,” in *Radiomics and Radiogenomics in Neuro-oncology* (H. Mohy-ud Din and S. Rathore, eds.), vol. 11991, pp. 36–43, Cham: Springer International Publishing, 2020. Series Title: Lecture Notes in Computer Science.
- [83] D. Zhang, J. Luan, B. Liu, A. Yang, K. Lv, P. Hu, X. Han, H. Yu, A. Shmuel, G. Ma, and C. Zhang, “Comparison of MRI radiomics-based machine learning survival models in predicting prognosis of glioblastoma multiforme,” *Frontiers in Medicine*, vol. 10, p. 1271687, Nov. 2023.
- [84] G. Hajianfar, A. Haddadi Avval, S. A. Hosseini, M. Nazari, M. Oveisi, I. Shiri, and H. Zaidi, “Time-to-event overall survival prediction in glioblastoma multiforme patients using magnetic resonance imaging radiomics,” *La radiologia medica*, vol. 128, pp. 1521–1534, Dec. 2023.
- [85] A. Duman, X. Sun, S. Thomas, J. R. Powell, and E. Spezi, “Reproducible and Interpretable Machine Learning-Based Radiomic Analysis for Overall Survival Prediction in Glioblastoma Multiforme,” *Cancers*, vol. 16, p. 3351, Sept. 2024.
- [86] A. Carré, G. Klausner, M. Edjlali, M. Lerousseau, J. Briend-Diop, R. Sun, S. Ammari, S. Reuzé, E. Alvarez Andres, T. Estienne, S. Niyoteka, E. Battistella, M. Vakalopoulou, F. Dhermain, N. Paragios, E. Deutsch, C. Oppenheim, J. Pallud, and C. Robert, “Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics,” *Scientific Reports*, vol. 10, p. 12340, July 2020.
- [87] H. Moradmand, S. M. R. Aghamiri, and R. Ghaderi, “Impact of image preprocessing methods on reproducibility of radiomic features in multimodal magnetic resonance imaging in glioblastoma,” *Journal of Applied Clinical Medical Physics*, vol. 21, pp. 179–190, Jan. 2020.
- [88] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee, “N4itk: Improved n3 bias correction,” *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.
- [89] A. Zwanenburg, M. Vallières, M. A. Abdalah, H. J. W. L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R. J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G. J. R. Cook, C. Davatzikos, A. Depeursinge, M.-C. Desseroit, N. Dinapoli, C. V. Dinh, S. Echegaray, I. El Naqa, A. Y. Fedorov, R. Gatta, R. J. Gillies, V. Goh, M. Götz, M. Guckenberger, S. M. Ha, M. Hatt, F. Isensee, P. Lambin, S. Leger, R. T. Leijenaar, J. Lenkowicz, F. Lippert, A. Losnegård, K. H. Maier-Hein, O. Morin, H. Müller, S. Napel, C. Nioche, F. Orlhac, S. Pati, E. A. Pfaehler, A. Rahmim, A. U. Rao, J. Scherer, M. M. Siddique, N. M. Sijtsema, J. Socarras Fernandez, E. Spezi, R. J.

- Steenbakkens, S. Tanadini-Lang, D. Thorwarth, E. G. Troost, T. Upadhaya, V. Valentini, L. V. van Dijk, J. van Griethuysen, F. H. van Velden, P. Whybra, C. Richter, and S. Löck, “The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping,” *Radiology*, vol. 295, no. 2, pp. 328–338, 2020. PMID: 32154773.
- [90] B. Taha, D. Boley, J. Sun, and C. C. Chen, “State of Radiomics in Glioblastoma,” *Neurosurgery*, vol. 89, pp. 177–184, Aug. 2021.
- [91] S. Bijari, A. Jahanbakhshi, P. Hajishafiezharamini, and P. Abdolmaleki, “Differentiating Glioblastoma Multiforme from Brain Metastases Using Multidimensional Radiomics Features Derived from MRI and Multiple Machine Learning Models,” *BioMed Research International*, vol. 2022, p. 2016006, Jan. 2022.
- [92] N. Q. K. Le, T. N. K. Hung, D. T. Do, L. H. T. Lam, L. H. Dang, and T.-T. Huynh, “Radiomics-based machine learning model for efficiently classifying transcriptome subtypes in glioblastoma patients from MRI,” *Computers in Biology and Medicine*, vol. 132, p. 104320, May 2021.
- [93] D. Homola, “Borutapy: Python implementation of the boruta feature selection algorithm.” https://github.com/scikit-learn-contrib/boruta_py, 2026. GitHub repository, Accessed: 2026-05-20.
- [94] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. Springer Series in Statistics, New York, NY: Springer New York, 2009.
- [95] S. Pölsterl, “scikit-survival: A library for time-to-event analysis built on top of scikit-learn,” *Journal of Machine Learning Research*, vol. 21, no. 212, pp. 1–6, 2020.
- [96] C. Davidson-Pilon, “lifelines, survival analysis in python,” Oct. 2024.
- [97] B. B. Sierra, C. McLean, P. S. Hall, and C. A. Vallejos, “The c-index multiverse,” 2025.
- [98] J. Y. Verbakel, E. W. Steyerberg, H. Uno, B. De Cock, L. Wynants, G. S. Collins, and B. Van Calster, “ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models,” *Journal of Clinical Epidemiology*, vol. 126, pp. 207–216, Oct. 2020.
- [99] J. Deng, M. E. Elghobashy, K. Zang, S. K. Patel, E. Guo, and K. Heybati, “So You’ve Got a High AUC, Now What? An Overview of Important Considerations when Bringing Machine-Learning Models from Computer to Bedside,” *Medical Decision Making*, vol. 45, pp. 640–653, Aug. 2025.
- [100] M. Assel, D. D. Sjöberg, and A. J. Vickers, “The Brier score does not evaluate the clinical utility of diagnostic tests or prediction models,” *Diagnostic and Prognostic Research*, vol. 1, p. 19, Dec. 2017.
- [101] I. Averbuch, A. Bareket-Samish, D. A. Goldstein, S. Eizenstein, G. Markel, E. Rosenbaum, D. Limon, D. Bomze, E. B. Ludmir, and T. Meirson, “Challenges in Interpreting Survival Metrics in Clinical Trials: The Utility of Restricted Mean Survival Analyses,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 122, pp. 224–233, June 2025.

- [102] N. F. Brown, D. Ottaviani, J. Tazare, J. Gregson, N. Kitchen, S. Brandner, N. Fersht, and P. Mulholland, “Survival Outcomes and Prognostic Factors in Glioblastoma,” *Cancers*, vol. 14, p. 3161, June 2022.
- [103] J. Liu, C. Li, Y. Wang, P. Ji, S. Guo, Y. Zhai, N. Wang, M. Lou, M. Xu, M. Chao, Y. Jiao, W. Zhao, F. Feng, Y. Qu, S. Ge, and L. Wang, “Prognostic and Predictive Factors in Elderly Patients With Glioblastoma: A Single-Center Retrospective Study,” *Frontiers in Aging Neuroscience*, vol. 13, p. 777962, Jan. 2022.
- [104] J. W. Cozzens, N. B. Drewes, K. R. Delfino, K. L. Chin, D. V. Amin, B. C. Lokaitis, J. A. Espinosa, B. A. Jones, L. J. Acakpo-Satchivi, H. Dayoub, M. B. Frankel, E. Agamah, K. Rao, C. M. Bradbury, and J. Gao, “Age is not a significant predictor of survival in patients with IDH-wildtype glioblastoma that undergo gross total resection and complete adjuvant chemoradiation,” *Frontiers in Oncology*, vol. 15, p. 1657867, Sept. 2025.
- [105] S. W. Choi, H.-H. Cho, H. Koo, K. R. Cho, K.-H. Nenning, G. Langs, J. Furtner, B. Baumann, A. Woehrer, H. J. Cho, J. K. Sa, D.-S. Kong, H. J. Seol, J.-I. Lee, D.-H. Nam, and H. Park, “Multi-Habitat Radiomics Unravels Distinct Phenotypic Subtypes of Glioblastoma with Clinical and Genomic Significance,” *Cancers*, vol. 12, p. 1707, June 2020.
- [106] M. Yu, J. Liu, W. Zhou, X. Gu, and S. Yu, “MRI radiomics based on machine learning in high-grade gliomas as a promising tool for prediction of CD44 expression and overall survival,” *Scientific Reports*, vol. 15, p. 7433, Mar. 2025.
- [107] Z. Qian, Y. Li, Z. Sun, X. Fan, K. Xu, K. Wang, S. Li, Z. Zhang, T. Jiang, X. Liu, and Y. Wang, “Radiogenomics of lower-grade gliomas: a radiomic signature as a biological surrogate for survival prediction,” *Aging*, vol. 10, pp. 2884–2899, Oct. 2018.
- [108] J. Pérez-Beteta, D. Molina-García, J. A. Ortiz-Alhambra, A. Fernández-Romero, B. Luque, E. Arregui, M. Calvo, J. M. Borrás, B. Meléndez, Rodríguez De Lope, R. Moreno De La Presa, L. Iglesias Bayo, J. A. Barcia, J. Martino, C. Velásquez, B. Asenjo, M. Benavides, I. Herruzo, A. Revert, E. Arana, and V. M. Pérez-García, “Tumor Surface Regularity at MR Imaging Predicts Survival and Response to Surgery in Patients with Glioblastoma,” *Radiology*, vol. 288, pp. 218–225, July 2018.
- [109] P. Lin, J.-S. Pang, Y.-D. Lin, Q. Qin, J.-Y. Lv, G.-Q. Zhou, T.-M. Tan, W.-J. Mo, and G. Chen, “Tumour surface regularity predicts survival and benefit from gross total resection in IDH-wildtype glioblastoma patients,” *Insights into Imaging*, vol. 16, p. 42, Feb. 2025.
- [110] S. Bakas, C. Sako, H. Akbari, M. Bilello, A. Sotiras, G. Shukla, J. D. Rudie, N. F. Santamaría, A. F. Kazerooni, S. Pati, S. Rathore, E. Mamourian, S. M. Ha, W. Parker, J. Doshi, U. Baid, M. Bergman, Z. A. Binder, R. Verma, R. A. Lustig, A. S. Desai, S. J. Bagley, Z. Mourelatos, J. Morrisette, C. D. Watt, S. Brem, R. L. Wolf, E. R. Melhem, M. P. Nasrallah, S. Mohan, D. M. O’Rourke, and C. Davatzikos, “The University of Pennsylvania glioblastoma (UPenn-GBM) cohort: advanced MRI, clinical, genomics, & radiomics,” *Scientific Data*, vol. 9, p. 453, July 2022.

- [111] E. Calabrese, J. Villanueva-Meyer, J. Rudie, A. Rauschecker, U. Baid, S. Bakas, S. Cha, J. Mongan, and C. Hess, “The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM),” Apr. 2023.
- [112] S. Cepeda, S. García García, I. Arrese, F. Herrero, T. Escudero Caro, T. Zamora Martínez, and R. Sarabia, “The Río Hortega University Hospital Glioblastoma dataset: a comprehensive collection of preoperative, early postoperative and recurrence MRI scans (RHUH-GBM),” 2023.
- [113] A. Carré, E. Battistella, S. Niyoteka, R. Sun, E. Deutsch, and C. Robert, “AutoComBat: a generic method for harmonizing MRI-based radiomic features,” *Scientific Reports*, vol. 12, p. 12762, July 2022.

Chapter A

Appendix

A.1 Textures Features

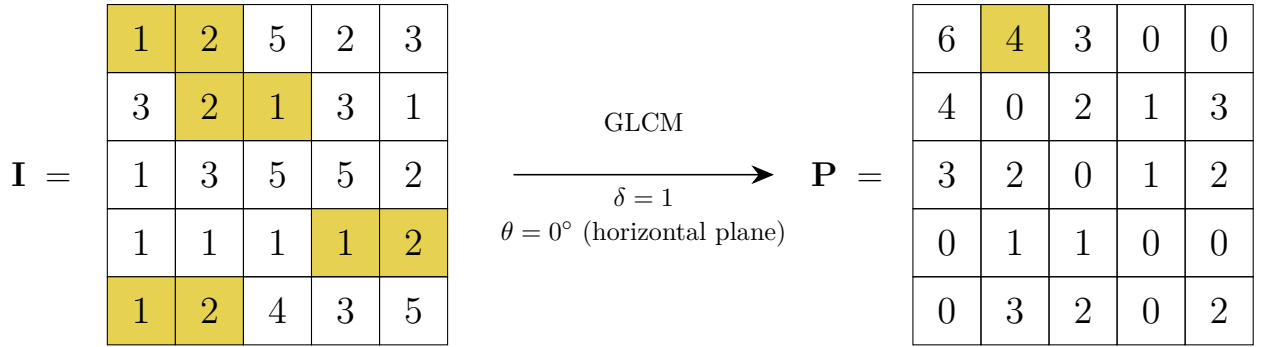


Figure A.1: 2D example of GLCM construction on a 5×5 image **I** with $N_g = 5$ gray levels, computed along the horizontal direction ($\theta = 0^\circ$) at distance $\delta = 1$. Each entry $P(i, j)$ counts the number of voxel pairs with gray levels (i, j) separated by δ voxels along θ , with symmetric counting (both (i, j) and (j, i) are accumulated). The four highlighted voxel pairs in **I** correspond to the combination $(1, 2) / (2, 1)$ and contribute to the highlighted entry $P(1, 2) = 4$. Example adapted from [32].

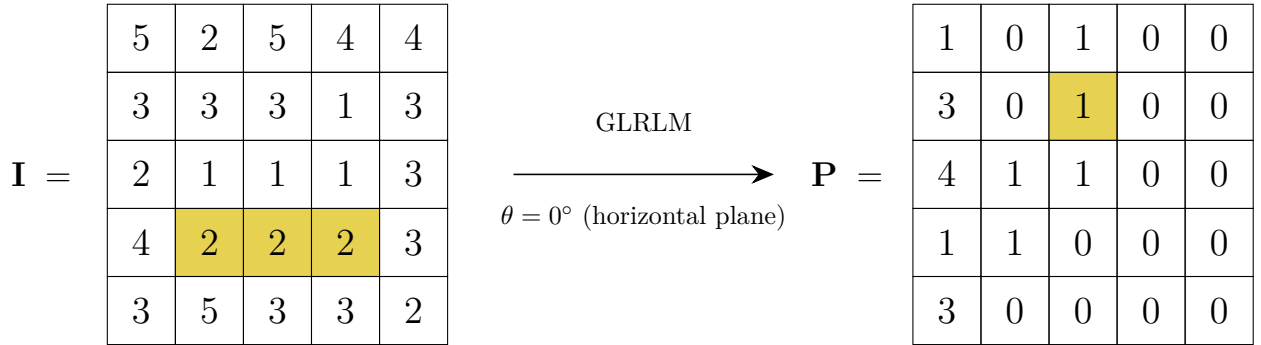


Figure A.2: 2D example of GLRLM construction on a 5×5 image **I**, computed along the horizontal direction ($\theta = 0^\circ$). Each entry $P(i, j)$ counts the number of runs of length j made of voxels with gray level i , where a run is a maximal sequence of consecutive voxels sharing the same gray level along θ . The highlighted run "2 2 2" of length 3 in the fourth row of **I** contributes to the highlighted entry $P(2, 3) = 1$. Example adapted from [32].

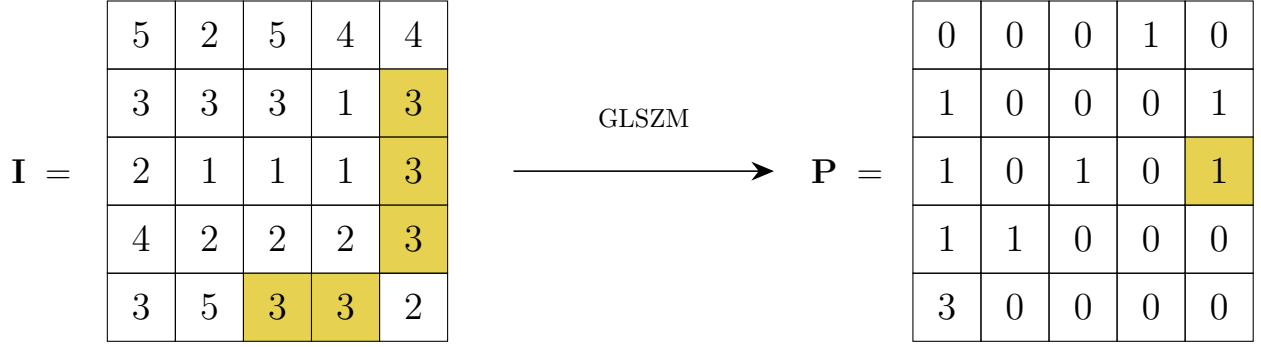


Figure A.3: 2D example of GLSZM construction on the same 5×5 image $\mathbf{I}$ as Figure A.2. Unlike the GLRLM, the GLSZM is direction-independent: each entry $P(i, j)$ counts the number of connected zones of j voxels sharing gray level i , where connectivity is taken in the 8-neighbourhood in 2D (26-neighbourhood in 3D). The highlighted zone in $\mathbf{I}$ is a 5-voxel cluster of gray level 3 connected via diagonal adjacency, which contributes to the highlighted entry $P(3, 5) = 1$. Example adapted from [32].

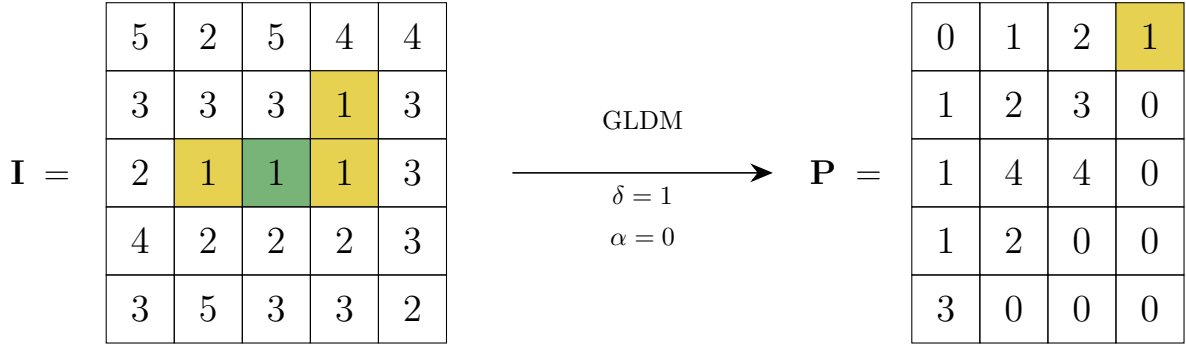


Figure A.4: 2D example of GLDM construction on a 5×5 image $\mathbf{I}$, with neighbourhood radius $\delta = 1$ (8-neighbourhood) and dependence tolerance $\alpha = 0$. Each entry $P(i, j)$ counts the number of voxels with gray level i that have exactly j dependent neighbours, where a neighbour k is dependent on the central voxel c if $|I(k) - I(c)| \leq \alpha$. With $\alpha = 0$, dependence reduces to strict equality of gray levels. The matrix columns correspond to $j \in \{0, 1, 2, 3\}$. In $\mathbf{I}$, the central voxel of gray level 1 (in green) has three dependent neighbours of gray level 1 (in yellow), contributing to the highlighted entry $P(1, 3) = 1$. Example adapted from [32].

$\mathbf{I} =$	1	2	5	2
	3	5	1	3
	1	3	5	5
	3	1	1	1

$\xrightarrow{\text{NGTDM}}$

$\mathbf{P} =$	i	n_i	p_i	s_i
	1	6	0.375	13.35
	2	2	0.125	2.00
	3	4	0.25	3.03
	4	0	0.00	0.00
	5	4	0.25	10.075

Figure A.5: 2D example of NGTDM construction on a 4×4 image $\mathbf{I}$ with $N_g = 5$ gray levels (level 4 is absent from $\mathbf{I}$), computed with neighbourhood radius $\delta = 1$ (8-neighbourhood). The NGTDM is a one-dimensional table indexed by gray level i and stores, for each i :

- n_i , the number of voxels with gray level i ;
- $p_i = n_i/N$, the empirical probability of gray level i , with N the total number of voxels used;
- $s_i = \sum_{k: I(k)=i} |i - \bar{A}_k|$, the cumulative absolute difference between gray level i and the mean gray level $\bar{A}_k$ of the neighbours of voxel k , summed over all voxels of gray level i .

Border voxels are included with their truncated neighbourhood. The row $i = 4$ contains zeros because no voxel of $\mathbf{I}$ has gray level 4. Example adapted from [32].

A.2 All features extracted from PyRadiomics

Table A.1: Radiomic features extracted by PyRadiomics, sorted by class. The * symbol indicates features that are disabled by default (and must be explicitly enabled using `enableFeaturesByName`).

Class	#	Feature
First Order Statistics — 19 features		
	1	Energy
	2	Total Energy
	3	Entropy
	4	Minimum
	5	10th Percentile
	6	90th Percentile
	7	Maximum
	8	Mean
	9	Median
	10	Interquartile Range
	11	Range
	12	Mean Absolute Deviation (MAD)
	13	Robust Mean Absolute Deviation (rMAD)
	14	Root Mean Squared (RMS)
	15	Standard Deviation*
	16	Skewness
	17	Kurtosis
	18	Variance
	19	Uniformity
Shape Features (3D) — 16 features		
	1	Mesh Volume
	2	Voxel Volume
	3	Surface Area
	4	Surface Area to Volume Ratio
	5	Sphericity
	6	Compactness 1*
	7	Compactness 2*
	8	Spherical Disproportion*
	9	Maximum 3D Diameter
	10	Maximum 2D Diameter (Slice)
	11	Maximum 2D Diameter (Column)
	12	Maximum 2D Diameter (Row)
	13	Major Axis Length
	14	Minor Axis Length
	15	Least Axis Length
	16	Elongation
	17	Flatness

continued on the next page

Table A.1 – continued from the previous page

Class	#	Feature
Gray Level Co-occurrence Matrix (GLCM) — 24 features		
	1	Autocorrelation
	2	Joint Average
	3	Cluster Prominence
	4	Cluster Shade
	5	Cluster Tendency
	6	Contrast
	7	Correlation
	8	Difference Average
	9	Difference Entropy
	10	Difference Variance
	11	Joint Energy
	12	Joint Entropy
	13	Informational Measure of Correlation 1 (IMC1)
	14	Informational Measure of Correlation 2 (IMC2)
	15	Inverse Difference Moment (IDM)
	16	Maximal Correlation Coefficient (MCC)
	17	Inverse Difference Moment Normalized (IDMN)
	18	Inverse Difference (ID)
	19	Inverse Difference Normalized (IDN)
	20	Inverse Variance
	21	Maximum Probability
	22	Sum Average
	23	Sum Entropy
	24	Sum of Squares
Gray Level Run Length Matrix (GLRLM) — 16 features		
	1	Short Run Emphasis (SRE)
	2	Long Run Emphasis (LRE)
	3	Gray Level Non-Uniformity (GLN)
	4	Gray Level Non-Uniformity Normalized (GLNN)
	5	Run Length Non-Uniformity (RLN)
	6	Run Length Non-Uniformity Normalized (RLNN)
	7	Run Percentage (RP)
	8	Gray Level Variance (GLV)
	9	Run Variance (RV)
	10	Run Entropy (RE)
	11	Low Gray Level Run Emphasis (LGLRE)
	12	High Gray Level Run Emphasis (HGLRE)
	13	Short Run Low Gray Level Emphasis (SRLGLE)
	14	Short Run High Gray Level Emphasis (SRHGLE)
	15	Long Run Low Gray Level Emphasis (LRLGLE)
	16	Long Run High Gray Level Emphasis (LRHGLE)

continued on the next page

Table A.1 – continued from the previous page

Class	#	Feature
Gray Level Size Zone Matrix (GLSZM) — 16 features		
	1	Small Area Emphasis (SAE)
	2	Large Area Emphasis (LAE)
	3	Gray Level Non-Uniformity (GLN)
	4	Gray Level Non-Uniformity Normalized (GLNN)
	5	Size-Zone Non-Uniformity (SZN)
	6	Size-Zone Non-Uniformity Normalized (SZNN)
	7	Zone Percentage (ZP)
	8	Gray Level Variance (GLV)
	9	Zone Variance (ZV)
	10	Zone Entropy (ZE)
	11	Low Gray Level Zone Emphasis (LGLZE)
	12	High Gray Level Zone Emphasis (HGLZE)
	13	Small Area Low Gray Level Emphasis (SALGLE)
	14	Small Area High Gray Level Emphasis (SAHGLE)
	15	Large Area Low Gray Level Emphasis (LALGLE)
	16	Large Area High Gray Level Emphasis (LAHGLE)
Neighbouring Gray Tone Difference Matrix (NGTDM) — 5 features		
	1	Coarseness
	2	Contrast
	3	Busyness
	4	Complexity
	5	Strength
Gray Level Dependence Matrix (GLDM) — 14 features		
	1	Small Dependence Emphasis (SDE)
	2	Large Dependence Emphasis (LDE)
	3	Gray Level Non-Uniformity (GLN)
	4	Dependence Non-Uniformity (DN)
	5	Dependence Non-Uniformity Normalized (DNN)
	6	Gray Level Variance (GLV)
	7	Dependence Variance (DV)
	8	Dependence Entropy (DE)
	9	Low Gray Level Emphasis (LGLE)
	10	High Gray Level Emphasis (HGLE)
	11	Small Dependence Low Gray Level Emphasis (SDLGLE)
	12	Small Dependence High Gray Level Emphasis (SDHGLE)
	13	Large Dependence Low Gray Level Emphasis (LDLGLE)
	14	Large Dependence High Gray Level Emphasis (LDHGLE)

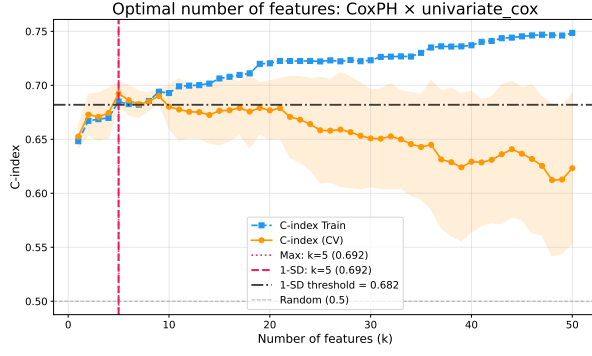
A.3 Results

Table A.2: Top 30 features selected by the univariate Cox approach. The C-index is the raw concordance index against overall survival; “Direction” indicates whether higher values of the feature are associated with increased (Risk) or decreased (Protective) hazard.

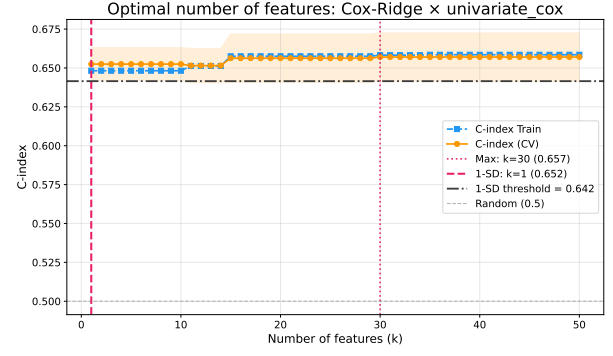
Rank	Feature	C-index	Direction
1	Age	0.648	Risk
2	t1__original_glszm_GrayLevelNonUniformity_ET	0.605	Risk
3	t1__original_ngtdm_Strength_NCR	0.595	Risk
4	t2__original_ngtdm_Busyness_ED	0.590	Risk
5	t2__wavelet-LHL_firstorder_Kurtosis_ET	0.588	Risk
6	t2__original_ngtdm_Coarseness_ED	0.587	Risk
7	t2__original_ngtdm_Busyness_ET	0.587	Risk
8	t2__original_ngtdm_Strength_ET	0.586	Risk
9	flair__original_glszm_GrayLevelNonUniformity_NCR	0.585	Risk
10	flair__original_glszm_GrayLevelNonUniformity_ET	0.581	Risk
11	t2__original_firstorder_10Percentile_ED	0.580	Risk
12	flair__wavelet-LHH_firstorder_Energy_ET	0.580	Risk
13	t1ce__wavelet-LLL_ngtdm_Strength_NCR	0.580	Risk
14	t2__original_glcmm_Correlation_ET	0.578	Risk
15	flair__wavelet-LHL_glszm_GrayLevelNonUniformityNormalized_NCR	0.577	Risk
16	flair__wavelet-LLH_firstorder_Energy_ET	0.575	Risk
17	t1__wavelet-HLH_firstorder_Entropy_ET	0.425	Protective
18	flair__wavelet-LHH_glszm_GrayLevelNonUniformityNormalized_NCR	0.575	Risk
19	flair__original_shape_Maximum2DDiameterColumn_ED	0.574	Risk
20	t2__original_glcmm_ClusterShade_ET	0.573	Risk
21	t1__original_ngtdm_Busyness_NCR	0.428	Protective
22	flair__original_glszm_GrayLevelNonUniformity_ED	0.572	Risk
23	flair__wavelet-LHL_glrmm_GrayLevelNonUniformityNormalized_ET	0.572	Risk
24	t2__wavelet-LLL_ngtdm_Strength_ET	0.571	Risk
25	t2__wavelet-HLH_glszm_GrayLevelNonUniformityNormalized_NCR	0.570	Risk
26	flair__original_shape_LeastAxisLength_NCR	0.570	Risk
27	flair__original_shape_MajorAxisLength_NCR	0.569	Risk
28	flair__wavelet-LHL_firstorder_Energy_ET	0.569	Risk
29	t2__original_glszm_ZoneEntropy_ET	0.568	Risk
30	t2__wavelet-HLL_firstorder_10Percentile_ET	0.568	Risk

Table A.3: Top 30 features selected by permutation importance. Importance is the mean decrease in model performance when the feature is randomly permuted.

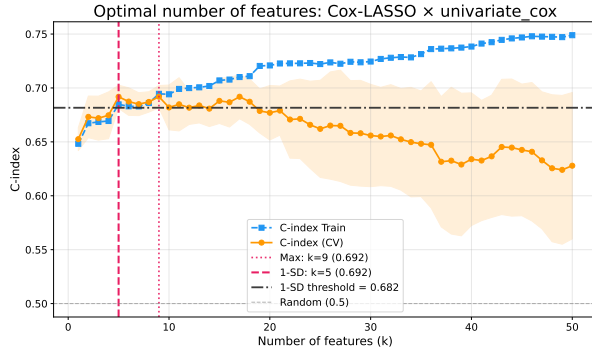
Rank	Feature	Importance
1	Age	6.823×10^{-3}
2	t1__wavelet-LHL_firstorder_Mean_ED	2.811×10^{-3}
3	flair__original_shape_Elongation_ET	2.035×10^{-3}
4	flair__original_glszm_GrayLevelNonUniformity_NCR	1.879×10^{-3}
5	flair__wavelet-LHH_glrlm_HighGrayLevelRunEmphasis_ET	1.869×10^{-3}
6	t1ce__original_firstorder_90Percentile_NCR	1.777×10^{-3}
7	t2__original_glcm_Correlation_ET	1.775×10^{-3}
8	flair__wavelet-LLH_glcm_ClusterProminence_NCR	1.692×10^{-3}
9	flair__original_shape_Flatness_ED	1.682×10^{-3}
10	t2__wavelet-HLH_glcm_Autocorrelation_ET	1.655×10^{-3}
11	t2__wavelet-LHH_firstorder_Entropy_ED	1.591×10^{-3}
12	t2__wavelet-LLH_glszm_ZoneEntropy_ET	1.582×10^{-3}
13	t2__wavelet-LHL_firstorder_Kurtosis_ET	1.561×10^{-3}
14	flair__wavelet-HHL_glszm_HighGrayLevelZoneEmphasis_ED	1.524×10^{-3}
15	t1ce__wavelet-HHL_firstorder_Mean_ET	1.508×10^{-3}
16	t1ce__wavelet-LHL_glcm_SumEntropy_ET	1.470×10^{-3}
17	t2__wavelet-LHL_firstorder_Skewness_ET	1.443×10^{-3}
18	t1ce__wavelet-LHH_firstorder_Kurtosis_NCR	1.370×10^{-3}
19	t1__wavelet-LHL_firstorder_Entropy_ED	1.300×10^{-3}
20	t1ce__original_glrlm_LongRunHighGrayLevelEmphasis_ED	1.256×10^{-3}
21	t1ce__original_firstorder_InterquartileRange_ED	1.225×10^{-3}
22	t2__wavelet-LHL_firstorder_Entropy_NCR	1.213×10^{-3}
23	t1ce__wavelet-LLH_firstorder_10Percentile_ET	1.192×10^{-3}
24	t2__wavelet-LHL_glszm_HighGrayLevelZoneEmphasis_ET	1.182×10^{-3}
25	t1ce__wavelet-HLL_glszm_SmallAreaHighGrayLevelEmphasis_ED	1.140×10^{-3}
26	t1ce__wavelet-HHL_gldm_DependenceEntropy_ET	1.111×10^{-3}
27	t1__wavelet-LHH_firstorder_Median_NCR	1.109×10^{-3}
28	flair__wavelet-HLH_glrlm_HighGrayLevelRunEmphasis_ED	1.107×10^{-3}
29	t2__wavelet-LHH_glszm_HighGrayLevelZoneEmphasis_NCR	1.090×10^{-3}
30	t1__wavelet-LHH_firstorder_Median_ET	1.076×10^{-3}



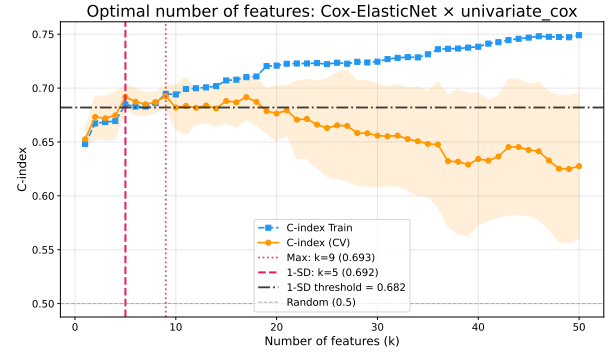
(a) CoxPH



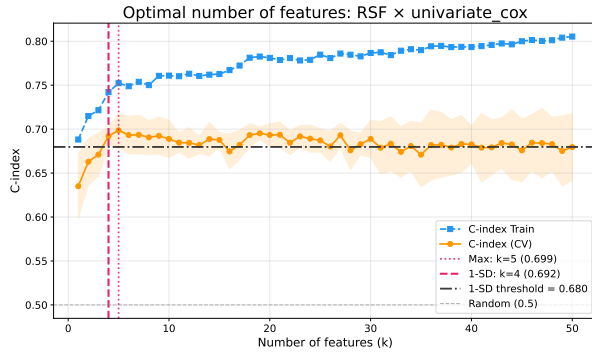
(b) Cox-Ridge



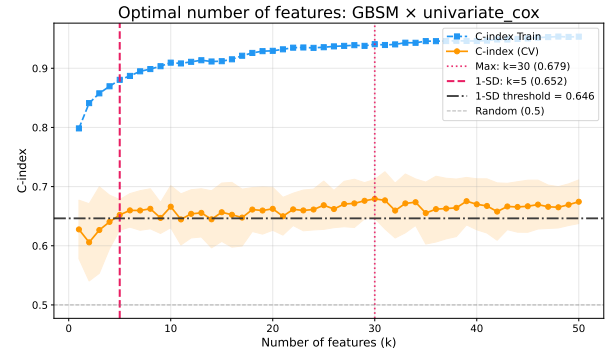
(c) Cox-LASSO



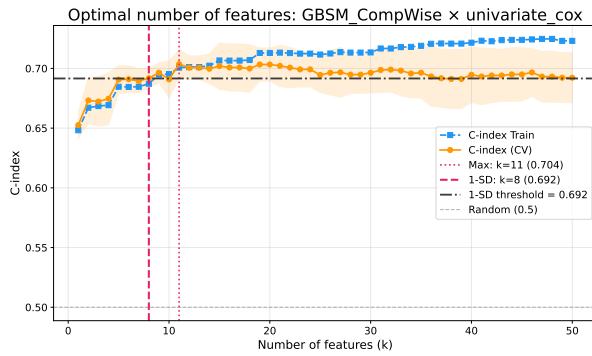
(d) Cox-ElasticNet



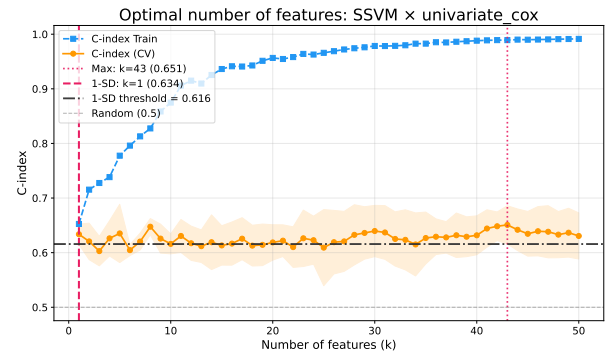
(e) RSF



(f) GBM

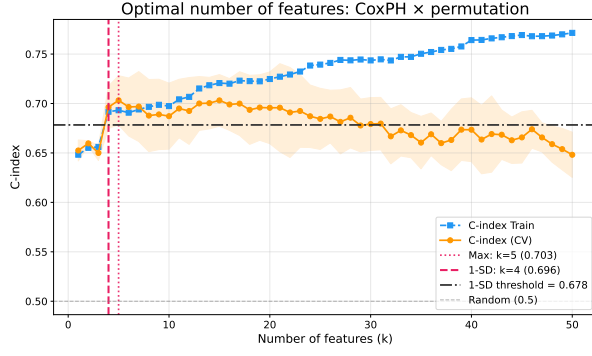


(g) GBM CompWise

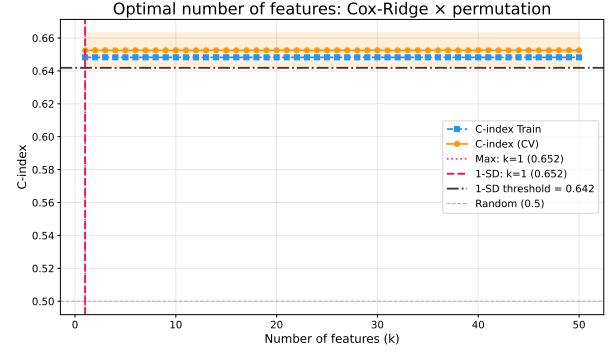


(h) SSVM

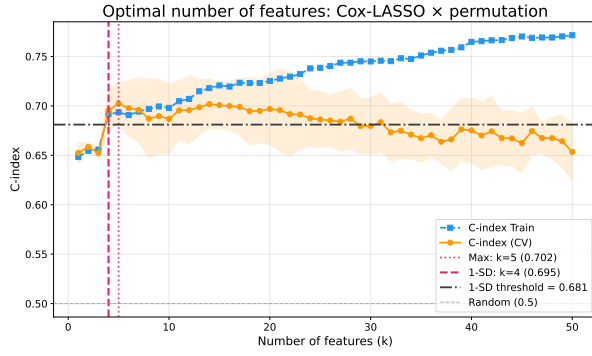
Figure A.6: Optimal k : univariate Cox method, per model.



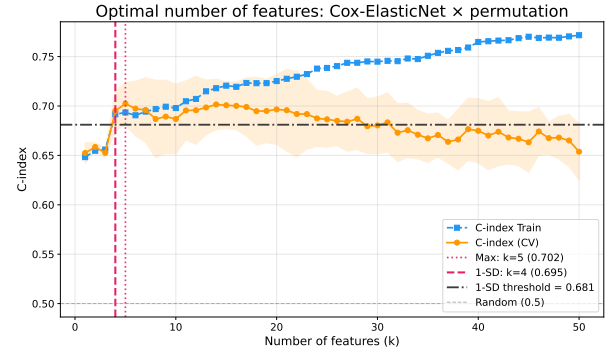
(a) CoxPH



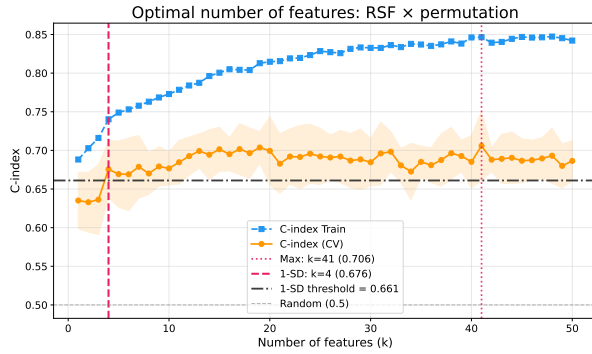
(b) Cox-Ridge



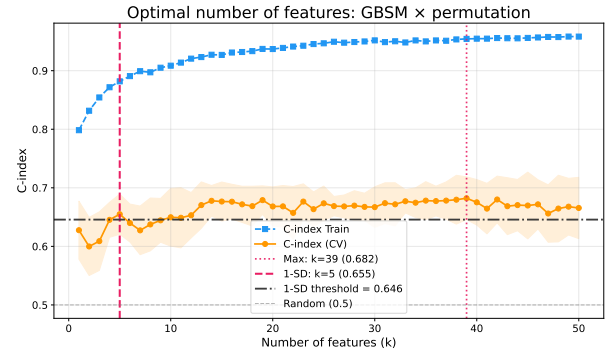
(c) Cox-LASSO



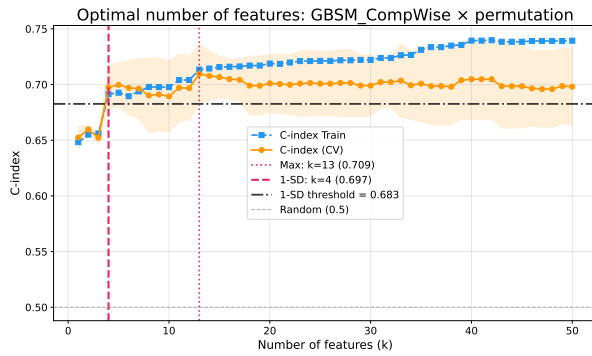
(d) Cox-ElasticNet



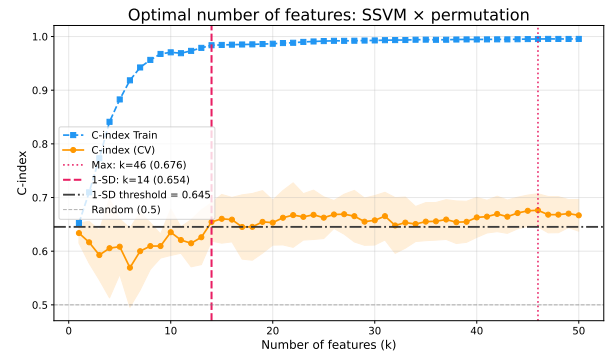
(e) RSF



(f) GBM



(g) GBM CompWise



(h) SSVM

Figure A.7: Optimal k : permutation method, per model.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
École polytechnique de Louvain

Rue Archimède, 1 bte L6.11.01, 1348 Louvain-la-Neuve, Belgique | www.uclouvain.be/epl