

Interpretable Machine Learning Classifiers for Brain Tumour Survival Prediction

Colleen E. Charlton^a, Michael Tin Chung Poon^{b,c,d,e}, Paul M. Brennan^{b,c,d}, Jacques D. Fleuriot^a

^a*Artificial Intelligence and its Applications Institute, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, UK*

^b*Cancer Research UK Brain Tumour Centre of Excellence, CRUK Edinburgh Centre, University of Edinburgh, Edinburgh, UK*

^c*Department of Clinical Neuroscience, Royal Infirmary of Edinburgh, 51 Little France Crescent, EH16 4SA, UK.*

^d*Translational Neurosurgery, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK*

^e*Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK*

Abstract

Prediction of survival in patients diagnosed with a brain tumour is challenging because of heterogeneous tumour behaviours and responses to treatment. Better estimations of prognosis would support treatment planning and patient support. Advances in machine learning have informed development of clinical predictive models, but their integration into clinical practice is almost non-existent. One reason for this is the lack of interpretability of models. In this paper, we use a novel brain tumour dataset to compare two interpretable rule list models against popular machine learning approaches for brain tumour survival prediction. All models are quantitatively evaluated using standard performance metrics. The rule lists are also qualitatively assessed for their interpretability and clinical utility. The interpretability of the “black box” machine learning models is evaluated using two post-hoc explanation techniques, LIME and SHAP. Our results show that the rule lists were only slightly outperformed by the black box models. We demonstrate that rule list algorithms produced simple decision lists that align

Email addresses: Colleen.Charlton@camh.ca (Colleen E. Charlton), michael.poon@ed.ac.uk (Michael Tin Chung Poon), paul.brennan@ed.ac.uk (Paul M. Brennan), jdf@ed.ac.uk (Jacques D. Fleuriot)

Current address for C.E. Charlton: Krembil Centre for Neuroinformatics, Centre for Addiction and Mental Health, Toronto, Ontario, Canada

with clinical expertise. By comparison, post-hoc interpretability methods applied to “black box” models may produce unreliable explanations of local model predictions. Model interpretability is essential for understanding differences in predictive performance and for integration into clinical practice.

Keywords: Bayesian rule lists, interpretable models, machine learning, brain cancer, survival

1. Introduction

Glioblastomas (GBM) are the most common malignant brain tumour and have the poorest outcomes. Average survival is 12-18 months and the 5-year survival rate is less than 5% [1]. Lower grade gliomas have an average survival of 7 years, but ultimately most progress to GBM and death [2]. An accurate prediction of prognosis for patients would inform treatment planning and patient support, but this is challenging. Various factors impact prognosis, but the precise contribution of each factor, and combination of factors to outcomes appears to vary between patients.

At the group level, basic statistical models are well-established in brain tumour survival analysis, but patient level survival prediction remains a challenge, possibly because of the well described biological heterogeneity of the disease and of treatment responses. Several survival studies have used the Cox proportional hazards (Cox PH) model [3], a popular survival model in statistics, to identify relevant clinical features for the construction of a brain tumour nomogram [4, 5, 6, 7] which graphically depict a statistical model to be used to estimate individualised cancer prognosis [8]. However, simple multivariable regression techniques are ineffective at identifying novel informative patterns from data [9], so machine learning (ML) approaches are increasingly being explored in brain tumour survival analyses [10, 11] (see also Kourou et al. [12] and Wang et al. [13] for general reviews). Such studies often use genetic or imaging data, with complex black box models to make prognostic predictions. In clinical practice this type of data is seldom, hence these models are of little use. Fulop et al.

[14] used a large clinical and molecular brain tumour dataset to predict 400-, 900- and greater than 900-day survival after surgery using several ML methods. The authors found that a neural network model performed the best with an accuracy of 59%. Furthermore, the authors used LIME [15] to understand the main drivers behind the wrong predictions and emphasized the importance of model interpretability for clinical decision-making. Senders et al. [16] recently used demographic, socioeconomic, clinical, and radiographic features for the creation of an online calculator for the prediction of glioblastoma survival. The authors compared 15 ML and statistical algorithms and an Accelerated Failure Time [17] algorithm was selected. However, a follow-up Letter to the Editor in the same journal [18] noted inconsistencies in the model calculations and highlighted the danger of non-healthcare professionals accessing this online resource which may provide misleading information. Although ML can be used to build predictive models with superior performance and generalisability [13, 19], the implementation of such models in a clinical setting can come with safety, legal and ethical considerations. In healthcare, there is a desire for interpretable and explainable AI/ML models to give end users (e.g. clinicians) the support that will allow them to accept or reject predictions, thus enabling them to make informed judgements when it comes to high-stake medical decisions (see Ahmad et al. [20] and Holzinger et al. [21] for reviews on interpretable ML in healthcare).

There is no all-purpose definition of interpretability since this is a subjective concept that is often domain-specific [22, 23]. One way to define interpretability is as the degree to which a human can understand the cause of a decision [24]. Thus a model M_1 may be considered more interpretable than a model M_2 if M_1 's decisions are easier to comprehend than those of M_2 [25]. Interpretability may be achieved by either using an intrinsically interpretable model whereby its simple structure allows end-users to understand feature relationships and final predictions, or by applying post-hoc explanation techniques to analyse and extract information from a trained model [26]. Most ML models are not originally designed to be interpretable and advancement in ML performance has led to the belief in a model's accuracy-interpretability trade-off [22]. However,

interpretability may be used as a tool to improve accuracy [25] and models with interpretability constraints have already been shown to perform on par with unconstrained models across several healthcare domains [27, 28, 29].

In this paper we explore the use of Bayesian Rule Lists (BRL) [30] and Falling Rule Lists (FRL) [31] as two types of intrinsically interpretable ML models and apply them to the prediction of patient survival using a novel brain tumour dataset. Both models combine pre-mined frequent patterns from the dataset into a decision list using Bayesian statistics [30]. The FRL model is an extension of the BRL algorithm which produces an ordered decision list whereby the estimated probability of success decreases down the list [31]. The BRL and FRL algorithms are compared to a baseline Cox PH model and popular black box models by looking at the prediction for brain tumour survival. To represent the class of black box methods, we chose random forest (RF) [32], logistic regression (LR) [33] and support vector machine (SVM) [34], each of which have varying degrees of interpretability. Finally, in an attempt to understand the decisions made by non-transparent models, post-hoc interpretability methods LIME (Local Interpretable Model-Agnostic Explanations) [15] and SHAP (SHapley Additive exPlanations) [35] are applied to the black box models. For a given instance (e.g. patient), the explanations produced by the post-hoc methods are compared between the three ML models.

The remainder of this paper is organised as follows: In Section 2 we introduce both the raw and final dataset and the required preprocessing techniques; Section 3 outlines model construction; in Section 4 we report the model’s quantitative and qualitative results; in Section 5 we analyse and discuss the results and Section 6 closes with final conclusions.

2. The Brain Tumour Data

2.1. The Raw Dataset

This paper explores an anonymised hospital-based brain tumour dataset collected from routine electronic healthcare data of patients who presented to

regional neuro-oncology teams in the UK with a brain tumour diagnosis. Initially, the dataset contained 1296 patient records and 225 predictor variables. A preliminary exploratory analysis of the data led to the removal of incomplete variables (features less than 60% filled) ($n=179$) and variables irrelevant to predicting outcomes (e.g. location of first imaging, clinician ordering CT (open access CT), contrast agent) ($n=14$), as well as the grouping of duplicated predictor variables (e.g. Symptom 1 and Symptom 1 - other) ($n=10$). Additionally, a number of patients were removed due to incomplete records (i.e. a patient is missing more than 60% of the reduced predictor variables) ($n=51$) or because of a lack of symptomatology information (i.e. a patient did not present with any symptoms or signs) ($n=227$). The remaining 1018 patients records contained 21 predictor variables and one dependent variable, namely patient survival in days.

The raw dataset contained significant heterogeneity with more than 30 different brain tumour types, the most common being glioblastoma ($n=540$), followed by metastasis ($n=198$), glioma ($n=186$) and meningioma ($n=174$). The minimum age of diagnosis is 16 years while the oldest is 97 years, and a median age of 61 years. There are an almost equal number of male and female patients (51% and 49%, respectively). 18% of patients had a previous history of cancer and 48% of patients presented with a co-morbidity, the most common being cardiovascular (15% of all patients). The most common location for a tumour was in the frontal lobe (34%) followed by the temporal lobe (22%). More than half of the patients (68%) had some type of surgery and 23% of patients underwent chemotherapy. This is inline with current treatment protocols where surgery is often the first line of treatment, followed by chemotherapy and concurrent radiotherapy for malignant tumours [36]. Finally, 23% of patients received no treatment, reflecting either the benign nature of the tumour, or conversely its advanced state or poor clinical condition of the patient.

Patient survival was measured in days from radiological diagnosis of brain tumour and 35% of patients were still alive at the time of dataset analysis. This is known as *censored data*, which is common in survival analysis, whereby the

value of an observation, in this case survival, is only partially known [37]. In the raw dataset, the largest survival time is 3964 days, or about 10 years, however only 11 patients have a survival time greater than 2000 days. At the time of data analysis, all patients with censored survival data had been alive for more than one year following diagnosis.

2.2. Preprocessing

Next, we review various pre-processing steps that were needed to get the narrowed-down dataset of 1018 patients into a state suitable for our analyses. Figure 1 illustrates the percentage of patient records complete for each feature.

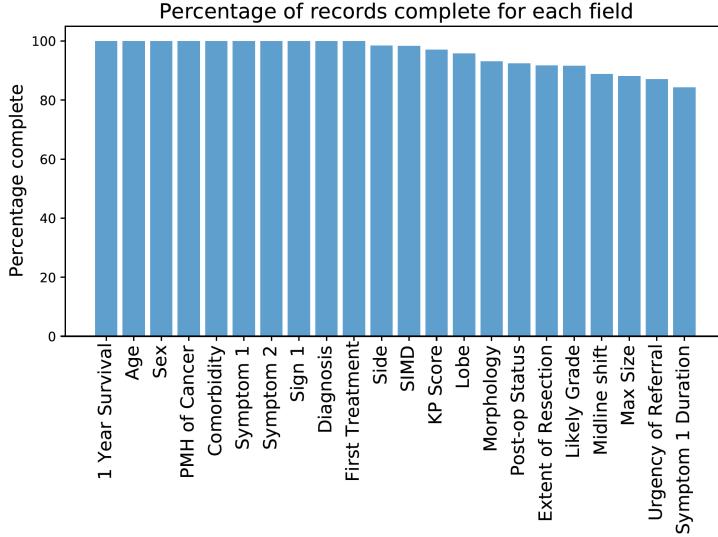


Figure 1: Percentage of records complete for each feature in the raw dataset. See the Appendix for feature descriptions.

2.2.1. Imputation

Given the small dataset size, missing data was managed through imputation rather than deletion. Treating all missing data the same would be a strong oversimplification as missing data can come from a variety of sources [38]. An entry for a feature may be absent, but this does not imply that the entry is

truly missing. For example, a patient may only present with one symptom thus leaving the remaining symptom features empty. This creates the appearance of missing data but in fact the empty entries are correctly missing. The variables to be imputed were all believed to be missing completely at random.

A number of imputation techniques were tried to determine which method was appropriate for each incomplete variable. In particular, a baseline mode-and mean-fill was used for categorical and continuous variables, respectively, and were compared to a k -nearest neighbours (k -NN) [39] and regression [40] imputation technique. For k -NN, the use of normalised and non-normalised features were both explored to account for differences in numerical values. The continuous variables in the dataset are constrained thus outliers were not a concern (e.g. the feature *Symptom 1 Duration* must fall between 0 and 52 weeks). For each variable with missing values, the imputation techniques were evaluated on the entire dataset using 10-fold cross-validation. The imputation of categorical and continuous variables was assessed using accuracy and the standard mean square error, respectively. The optimal imputation method for each variable was then implemented on the full dataset, whereby the missing variables were replaced with the model's predicted output. More information about the imputation methods are described in a forthcoming paper based on the following thesis [41].

2.2.2. Discretisation

Following imputation, all continuous variables (*Age*, *Symptom 1 Duration*, and *Maximum Tumour Size*) were discretised in order to support the association rule mining [42] employed by the BRL and FRL algorithms (see Section 3.3). Most rule-mining approaches make the (restrictive) assumption that all features are binary or categorical. As part of the discretisation process, we compared three methods: uniform binning, quantiles and k -means. Each feature was divided into groups ranging from size 2 to 12, increasing by increments of two. We chose the maximum to be 12 as it was a natural divider for the feature *Maximum Tumour Size*, which contained the largest range in values (0 - 120).

The accuracy of the discretised features was compared to the continuous feature, using the BRL model with default hyperparameters. Classification of patient survival greater than one year was evaluated on the entire dataset using 5-fold-cross validation.

2.2.3. Dealing with Feature Collinearity

Finally, in our dataset, two sets of features were found to be highly correlated: *Tumour Type* and *Likely Grade* as well as *First Treatment* and *Extent of Resection*. Tumour type refers to the kind of tumour a patient is diagnosed with, while likely grade is an indicator of how quickly a tumour is likely to grow or spread. A grade I/II tumour is benign (slow growing and unlikely to spread within the brain) and a grade III/IV tumour is malignant (fast growing and likely to spread within the brain). However, by definition, a glioblastoma is a grade IV glioma tumour. Thus to reduce collinearity in the data, Tumour Type and Likely Grade were combined into a single feature called *Diagnosis*. Tumour types were separated into benign (or low-grade) and malignant (or high-grade) categories (e.g. Glioma Benign and Glioma Malignant). Metastatic tumours encompass a mix of tumour types that originate elsewhere in the body. By the fact that they migrate to the brain, metastatic tumours are all aggressive and thus classified as malignant. The second set of features, *First Treatment* and *Extent of Resection* (EOR), also suffer from multicollinearity. EOR refers to the amount of cancerous cells removed during surgery (e.g. 90-99%) and is only relevant if the first treatment a patient receives is surgery, which is not always the case. Thus we integrated EOR information into First Treatment (e.g. Surgery 100%, Surgery 90-99%) to create a more informative feature - the feature name remained *First Treatment*.

2.2.4. The Final Dataset

Our final dataset contains 1018 patient records and 19 predictor variables including patient demographics (e.g. sex, age), medical history (e.g. history of cancer, comorbidity), symptom features (e.g. symptom types and duration),

radiological tumour analysis (e.g. diagnosis, morphology) and treatment details (e.g. first treatment, post-op performance status). Many of these prognostic factors have been well-documented in the literature including age, Karnofsky performance (KP) score, symptoms, morphology, diagnosis (i.e. tumour type, likely grade) and treatment [43, 44, 45, 46, 47]. Table 1 provides a detailed summary of the salient variables in the final dataset, including their descriptions, value and percentage of each value present in the dataset following imputation and discretisation (see Table 4 in the Appendix for a comprehensive overview of all predictor variables in the final dataset). We briefly discuss some of the salient features below and some final processing of the features.

Table 1: Overview of 9 salient dataset variables including their descriptions, value and percentage of each value present in the final dataset following imputation and discretisation.

Name	Description	Value	Proportion (%)
Age	the age of a patient	0-44	17.1
		45-54	18.7
		55-61	16.0
		62-67	15.8
		68-74	16.9
		75+	15.5
Sex	the sex of the patient	Male	50.5
		Female	49.5
Karnofsky	a common measure in oncology	100	37.6
Performance	to assess the functional state	90	28.6
Score	of a patient (see Figure 10 in (KP Score)	80	14.6
	Appendix)	≤ 70	19.2

Continued on next page

Table 1 – continued from previous page

Name	Description	Value	Proportion (%)
Symptom 1	the first symptom type a patient presented with (reported by the patient)	Focal Neurology Headache Fits/Faints/Falls Behavioural/Cognitive Other/Non-specific Non-specific Neurological	34.6 28.4 17.1 16.7 2.4 0.8
Sign 1	the first sign type a patient presented with (observed by the physician)	No Signs Neurological Cognitive Cranial Nerve Other Behavioural	42.7 36.2 15.0 5.0 0.8 0.3
Diagnosis (or Tumour Type)	the type of brain tumour a patient was diagnosed with	Glioma Malignant Metastasis Meningioma Benign Glioma Benign Rare Tumour Benign Lymphoma Malignant Meningioma Malignant Rare Tumour Malignant Hemangioblastoma Benign	46.5 19.0 13.6 7.1 4.7 4.1 2.3 1.5 1.2 Benign
Morphology	the histological classification of the tumour based on the cell types present	Heterogenous Homogenous	68.5 31.5
Post-operative Performance Status	a measure of a patient's level of functioning following surgery in terms of their ability for self-care, daily activity, and physical ability (see Table 7 in Appendix)	0 1 2 3 4 5 No Surgery	31.5 27.4 6.2 1.9 1.4 0.2 31.4

Continued on next page

Table 1 – continued from previous page

Name	Description	Value	Proportion (%)
First Treatment	the type of first cancer treatment	Surgery Removal 100%	16.0
		Surgery Removal 90-99%	24.4
		Surgery Removal 50-89%	6.4
		Surgery Removal <50%	4.9
		Biopsy	16.9
		Radiotherapy	5.5
		Chemootherapy	0.9
		Other (e.g. steroids)	2.5
		No Treatment	22.5

KP Score: This is a standard way of assessing a patient’s ability to perform everyday tasks [48]. The scale is a ‘gold standard’ in clinical oncology and is commonly used to determine a cancer patient’s expected tolerance to treatments (e.g. chemotherapy). The scores ranges from 0 (dead) to 100 (normal) and is scored in deciles, although the values are ordinal (see Table 10 in the Appendix for the original definition of the KP scores). This means that a value assigned to a patient is based on a ranking but the numerical value associated with this rank is not meaningful. Thus the difference between the values 70 and 90 is not equivalent to the difference between the values 40 and 60. Furthermore, the KP scale may be subject to bias [49]. A patient’s KP score is determined by clinicians, and when compiling a dataset this can result in inter-observer subjectivity [50, 51]. To reduce the bias associated with the KP score, values of 70 and below were aggregated due to their negative association with survival [52] (a KP score of 70 reflects someone who can ‘care for self, but who is unable to carry on normal activity or to do active work’). KP scores of 80 and above remained separate allowing for a more fine-grained analysis of the values associated with survival.

Symptom 1: A symptom is observed by the patient themselves (subjective) and is often what drives a patient to consult a physician. Symptom 1 refers

to the first symptom a patient presents with. The symptom data in the raw dataset had a high cardinality of 37 different symptom types, with many of these types pertaining to a small number of patients. Thus we decided to group symptom types into six overarching categories – e.g. Headache, Fits/Faints/Falls and Behavioural/Cognitive – based on work by Ozama et al. [53] to create a more homogeneous set of symptom types. An outline of the symptom groupings are summarised in Table 5 of the Appendix.

Sign 1: A *sign* is observed by a physician (objective). Sign 1 refers to the first sign a patient presents with. The sign data in the raw dataset also had a high cardinality (26 different types), thus the data was additionally grouped into six larger domains – e.g. neurological and cognitive– based on the advice of the consulting clinical experts (see Table 6 in the Appendix). Although all patients in the final reduced dataset presented with at least one symptom, 43% of patients did not present with any signs.

Diagnosis (Tumour Type): Brain tumours are broadly named based on the type of normal cell that they most resemble, and their location in the brain [54]. In the raw dataset the tumour types had a high cardinality with many entries referring to the same general tumour type (e.g. meningioma suprasellar and meningioma at cerebellopontine (CP) angle). Tumour types that appeared in less than 10 patients were grouped into a “Rare Tumour” category. Additionally, the tumour type may be benign (i.e. grade I/II), or malignant (i.e. grade III/IV). In the final dataset, the brain tumour types were reorganised based on type and malignancy (e.g. Glioma Benign, Glioma Malignant), and reduced to a cardinality of 9.

Morphology: Each tumour type can have a different sub-classification, which is represented as morphology. A tumour with heterogenous morphology contains diverse cell types with distinct molecular structure that may have different levels of sensitivity to treatment [55]. In comparison, a homogenous tumor contains the same or similar genetic or epigenetic characteristics [56]. Hence tumour morphology may be indicative of treatment response.

First Treatment: The first type of cancer treatment a patient receives is

based on the presumed type based on imaging, location of the tumour, and the patient’s overall health (e.g. KP score ≤ 70). Surgery, for example, may be the only treatment necessary depending on the grade of the tumour and extent of resection. Information on extent of resection was included in the first treatment types (e.g. Surgery 100%, Surgery 90-99%, Biopsy, etc.). Other treatment types include radiotherapy and chemotherapy.

3. Methods

3.1. Modelling Techniques

As the black box ML models (RF, LR and SVM) cannot directly handle categorical variables, one-hot encoding was used for all 19 features, which resulted in 94 different feature types. For ease in model feature comparison, one-hot encoded features were also used by the Cox PH model. Due to the small dataset size, nested cross validation [57] was implemented for hyperparameter tuning and model assessment. The inner loop, which is responsible for the model selection, used 3-fold cross-validation, and the outer loop, which is responsible for estimating model performance, used 5-fold cross-validation. Nested cross validation was repeated for three different random seeds resulting in a $3 \times 5 \times 3$ setup. For each model, the average accuracy [58], macro-F1 [59], and area under the receiver operating characteristic curve (AUROC) [60] were reported. All models performed binary classification and one year survival labels were created from the survival data. This resulted in a relatively even split of the dataset: 443 patients (44%) survived less than a year and 575 patients (56%) survived more than a year.

3.2. Cox PH Model

The Cox PH model is a popular regression model for survival analysis [3]. This semi-parametric model predicts the risk of an event occurring (e.g. death) as a function of time. The hazard function $\lambda(t)$ estimates the effect of observed variables x_k on a baseline hazard function $\lambda_0(t)$. The model can be expressed

as: $\lambda(t) = \lambda_0(t)\exp(\beta_1x_1 + \beta_2x_2 + \dots + \beta_kx_k)$, where β_1, \dots, β_k are the coefficients estimated from the data of the observed variables x_1, \dots, x_k . However, the Cox PH model makes the restrictive assumption that each predictor variable has a multiplicative effect on the baseline hazard function that remains constant over time. Thus nonlinear relationships between predictor variables cannot be modelled by this approach.

The Cox PH model was implemented using the Python-package lifelines [61] and served as an alternative statistical baseline for the ML models. Due to the one-hot encoding of the categorical variables, the resulting dataset incurred problems with high collinearity. That is, some of the independent variables were highly correlated which tends to inflate the estimated regression coefficients. Thus the first category per feature was dropped - i.e. for a variable with k categories, $k-1$ dummies remained. A penalizer was also added to the model which reduces the size of the coefficients during regression, thus controlling for high correlation and improving the stability of the estimates [62]. A search for the optimal penalty parameter between the set of values $\{10^{-3}, 10^{-2}, \dots, 10^1\}$ was performed and a value of 10^{-2} was selected.

The purpose of the Cox PH model is to predict survival time. Thus the model was trained using the continuous survival data and the predicted survival time was converted to binary one year survival labels that were then used to calculate the model's performance.

3.3. Rule Lists

Rule lists are a type of intrinsically interpretable model that produces a series of *if-then* rules, also known as decision (or production) rules, which are used to generate predictions. Each rule is composed of two different sets of items, a and b , also known as itemsets. A decision rule is an implication of the form $a \rightarrow b$ (or *if...then...*), where a is an antecedent that is followed by a consequent b . As an example, for the rule:

```
IF Diagnosis: Glioma Malignant AND First Treatment: Surgery 100%
THEN probability of Survival > 1 Year: 90%,
```

the antecedent a is “Diagnosis: Glioma Malignant AND First Treatment: Surgery 100%”, and the consequent b is “Survival > 1 Year”. If a rule (or set of rules) is satisfied, the model outputs a certain classification. Although the concept of decision rules is well known in AI, more recently, rule-based models have been constructed directly from data with the help of ML [63].

Instead of being crafted using domain knowledge, rules can be *learned* using frequent itemset mining [64, 65], which looks for common patterns in the data. Itemsets, consisting of these frequent patterns, are used to construct the association rules that are normally subject to constraints on minimum support and confidence [64], although other measures such as lift [66] are possible. In particular,

$$\text{Support} = \frac{\text{freq}(a, b)}{N}$$

refers to how frequently the itemsets appear in the data, where $\text{freq}(a, b)$ is the frequency of the itemsets containing items a and b , and N is the number of observations in the dataset. If the support threshold it is too large the algorithm may fail at finding the true patterns in the dataset, whereas a small minimum support may generate an excess amount of association rules that is not fit for effective use. Confidence, for its part, is defined as:

$$\text{Confidence} = \frac{\text{freq}(a, b)}{\text{freq}(a)}$$

and is the frequency of itemsets that contain a which also contain b , or how often a rule is found to be true [67]. Consequently, the performance of these rule mining algorithms is dependent on user-specified thresholds. For both the BRL and FRL models, rules were mined using the default minimum support threshold of 10% and confidence threshold of 80% that is commonly used in the literature. Different thresholds on minimum support and confidence were explored but model performance did not improve thus the default parameters were retained.

Finally, the BRL and FRL algorithms contain additional hyperparameters to specify the maximum rule cardinality and expected rule list length. The maximum rule cardinality refers to the length of the rule (or itemset). This

value is typically set to 2 or 3 as it may be harder to reconcile a combination of features as being clinically logical when interpreting high cardinality rules. The prior expected list length denotes the expected number of rules in the list (excluding the null rule).

3.3.1. Bayesian Rule Lists

BRLs [30] are used for classification problems where the goal is to learn $P(Y = 1|X)$. Y is binary, and in the case of predicting brain tumour survival greater than a year, $Y = 1$ would indicate survival greater than a year and X would represent a patient’s features. The conditional probability distribution is represented as a decision list consisting of a series of decision rules.

Frequent patterns (or itemsets) are first identified from the dataset using the rule mining algorithm FP-Growth [67]. Following association rule mining, the BRL algorithm creates a posterior distribution over rule lists, given the observed data and prior assumptions (i.e., max rule cardinality and list length). These priors are used to specify rule cardinality and rule list length. Using a generative model, an initial decision list is selected and iteratively modified using Markov chain Monte Carlo sampling [68] to generate many samples of decision lists from the posterior distribution (see Letham et al., [30] for technical details). This procedure ensures the production of a variety of lists that are not dependent on one initial decision list.

Given this posterior distribution of decision lists, new observations are classified using a point estimate (a single decision list) or the posterior predictive distribution (multiple decision lists). The point estimate is chosen as the list with the highest posterior probability from all the samples with posterior mean list length and posterior mean average rule cardinality. This estimate is called a *BRL-point* [30].

3.3.2. Falling Rule Lists

A FRL is an ordered decision list whereby the estimated probability of success, or $P(Y = 1 | X)$, monotonically decreases down the list [31]. Analogous to

BRL, pre-mined rules are first extracted from the data, then Bayesian modelling is used to produce a decision list (see Wang and Rudin [31] for mathematical details). To approximate the FRL, Monte Carlo sampling from the posterior distribution is required to generate an initial decision list. To enforce FRL monotonicity constraints, a combination of Gibbs and Metropolis-Hastings sampling [69] is used. Unlike BRL, following the production of an initial decision list, instead of yielding many sample lists, a point estimate is found using simulated annealing [70].

3.4. ML Algorithms

Three ML classifiers were exploited to predict the 1-year survival of patients. An RF classifier [32], which uses a multitude of decision trees [71] for classification, was implemented. The number of trees were selected using grid search with 3-fold cross validation. A LR classifier [33] with L2 regularisation [72] was also employed. The regularisation parameter C_{LR} was selected from $\{2^{-4}, 2^{-3}, 2^{-2}, \dots, 2^4\}$. Finally, an SVM classifier [34] with a radial basis function kernel [73] was used. The regularisation parameter C_{SVM} was selected from $\{2^{-4}, 2^{-3}, 2^{-2}, \dots, 2^4\}$. Both LR and SVM penalty parameters were chosen using a similar range of values to the original BRL paper [30]. All models were implemented using the scikit-learn package [74] and represent a group of relatively uninterpretable ML models.

4. Results

4.1. Model Evaluation

The mean classification performance of the six modelling approaches on the brain tumour dataset are summarised in Table 2. The ROC curves for the rule lists and ML models are illustrated in Figure 2. The mean ROC curves are in bold and the standard deviation is shown by the shaded region. The AUROC could not be directly computed for the Cox PH model because it requires the probability estimates for each class which the Cox PH model does

not provide. The baseline Cox PH model was outperformed by all models and FRL outperformed BRL. The rule lists were marginally outperformed by the three ML models, with SVM performing best. Notably, FRL performance was within one standard deviation of the mean performance of SVM and BRL performance was within two standard deviations. Thus the loss in BRL and FRL performance was minimal and may be mitigated by the rule list’s added level of interpretability.

	Cox	BRL	FRL	RF	LR	SVM
Accuracy	.711 (.045)	.758 (.028)	.782 (.021)	.807 (.024)	.805 (.030)	.809 (.030)
F1	.698 (.045)	.754 (.031)	.780 (.021)	.804 (.025)	.802 (.030)	.807 (.030)
AUROC	.711 (.052)	.759 (.033)	.784 (.021)	.805 (.025)	.804 (.029)	.808 (.029)

Table 2: Performance metrics were assessed using nested cross-validation for three different random seeds. Mean and, in parenthesis, the standard deviation for 15 models are given. The macro-averaged F1 score is reported.

Figure 3 and 4 show point-estimates for the BRL and FRL models obtained from one of the cross validation folds. For both types of rule lists, once a patient has satisfied a rule they will not be taken into account by the cases further down the list. The final rule in the list will only consider the subset of patients that were not classified by the previous ones. The shorter decision list produced by the FRL algorithm is likely due to the model’s monotonicity constraints. All rules that favour survival of less than a year are summarised into one final rule by the FRL algorithm. Comparatively, the BRL model does not follow any monotonicity constraints and so rules that favour both survival less than and greater than a year are included.

The BRL point-estimate (Figure 3) indicates that if a patient has a benign meningioma they will likely survive more than a year, however, if a patient receives no treatment and has a tumour with heterogeneous morphology, they will likely not survive more than a year. On the other hand, the FRL point-estimate (Figure 4) indicates that a tumour with homogenous morphology and a KP score of 100 are positive prognostic factors. Additionally, if a patient has

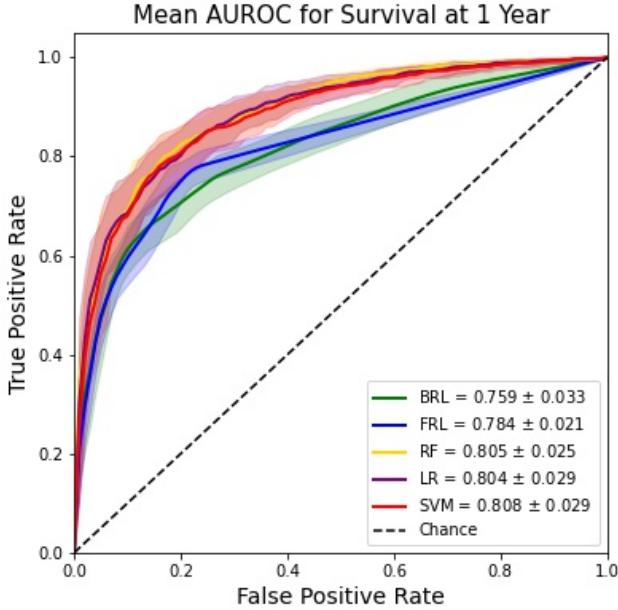


Figure 2: Mean AUROC and standard deviation for rule list and ML models are illustrated. The black dashed line represents a random classifier that performs no better than chance.

a 100% tumour resection and no previous history of cancer they are likely to survive greater than a year. By reviewing only a few rules from both lists, it is evident that features such as diagnosis, first treatment type, tumour morphology and KP score are significant survival predictors. Additionally, age of diagnosis and post-op status are repeated throughout both rule lists.

The BRL and FRL point-estimates from the four other folds for a specific random seed are given in Appendix B, for a total of five BRL and five FRL estimates. Despite the different point-estimates, there is significant overlap in the rules. Due to the iterative construction of BRLs, there may be multiple equally good rule lists produced and it is not clear which will be returned by the model [30]. The FRL model yields a single point-estimate using simulated annealing and only uses rules that favour survival greater than a year thus limiting the variability in its final decision list. The production of multiple high

```

IF Diagnosis : Meningioma Benign THEN probability of Survival > 1 year: 93.7%
(88.5%-97.4%)
ELSE IF First Treatment : None AND Morphology : Heterogeneous THEN probability of
Survival > 1 year: 7.8% (3.7%-13.4%)
ELSE IF KP Score : 80 AND Morphology : Heterogeneous THEN probability of Survival > 1
year: 33.3% (23.2%-44.3%)
ELSE IF Diagnosis : Glioma Benign THEN probability of Survival > 1 year: 98.3%
(93.8%-100.0%)
ELSE IF Age : (67.0, 74.0] AND Comorbidity : Yes THEN probability of Survival > 1 year:
8.7% (2.5%-18.3%)
ELSE IF Diagnosis : Metastasis AND Midline shift : 0 THEN probability of Survival > 1 year:
31.6% (18.0%-47.0%)
ELSE IF First Treatment : Surgery Removal 100% THEN probability of Survival > 1 year:
94.5% (87.3%-98.8%)
ELSE IF Age : (0.0, 44.0] THEN probability of Survival > 1 year: 89.9% (81.8%-95.8%)
ELSE IF KP Score : <=70 THEN probability of Survival > 1 year: 17.2% (8.7%-27.9%)
ELSE IF Midline shift : 0 THEN probability of Survival > 1 year: 84.1% (74.6%-91.6%)
ELSE IF Post-op Status : 0.0 THEN probability of Survival > 1 year: 70.8% (59.2%-81.1%)
ELSE probability of Survival > 1 year: 41.0% (30.4%-52.1%)

```

Figure 3: BRL-point estimate. The 95% credible interval is given in parentheses.

```

IF Morphology: Homogeneous AND KP Score: 100 THEN probability of Survival > 1 year is
98.08%, Support: 156
ELSE IF First Treatment: Surgery Removal 100% AND PMH of Cancer: No THEN probability of
Survival > 1 year is 92.86%, Support: 56
ELSE IF Age: (0.0, 44.0] THEN probability of Survival > 1 year is 86.44%, Support: 59
ELSE IF Diagnosis: Meningioma Benign THEN probability of Survival > 1 year is 83.33%,
Support: 30
ELSE IF Post-op Status: 0.0 THEN probability of Survival > 1 year is 66.07%, Support: 112
ELSE probability of Survival > 1 year is 26.12%, Support: 402

```

Figure 4: FRL-point estimate. The support indicates the number of patients classified by that rule.

performing rule lists may be beneficial, as additional information on feature relationships are revealed.

4.2. Model Interpretability

Although the algorithms do not provide the same level of interpretability, the weighting of features at a global model level and local prediction level can be reasonably compared. Sequential feature selection was first performed on the dataset to assess feature significance. The final models interpretability was evaluated using feature importance (Cox PH, RF), post-hoc methods LIME and SHAP (RF, LR, SVM) and qualitative assessment (BRL, FRL). These aspects are discussed next.

4.2.1. Feature Selection

Typically, the purpose of feature selection is to remove irrelevant features or noise from the data and improve computational efficiency. We used feature selection to assess which features were most pertinent to the data and to compare these results to the feature importance of the individual models.

Sequential feature selection from mlxtend¹ allows for a range of k -features to be specified and the feature combination that scores the best during cross validation is returned. Although SVM was the best performing model, the algorithm cannot be run in conjunction with the feature selector as a result of the kernel transformation, thus RF, the second best model, was run with the feature selector to provide a baseline for feature comparison. The RF default parameters were used and features were assessed using AUROC with 5-fold cross validation. Due to one-hot encoding, a total of 94 feature types were evaluated and the selector returned the 10 best features (see Table 3).

The selected features included diagnosis, age at diagnosis, SIMD score, KP score, morphology, post-op status and urgency of referral. Notably, many of these features have been previously mentioned in the literature as important prognostic variables [46, 47, 75, 76, 77], however SIMD, a social measure of deprivation in Scotland, is often associated with survival in a population study setting rather than a hospital-based setting. Nonetheless, an association be-

¹http://rasbt.github.io/mlxtend/use_guide/feature_selection/SequentialFeatureSelector/

tween socio-economic status and cancer survival has long been demonstrated in research [78]. Despite the UK’s universal healthcare, several cancer studies have linked deprivation with poorer survival outcome [79, 80, 81]. In our dataset, 62% of patients with a SIMD score of 4 or 5 (least deprived) survived greater than a year compared to 53% of patients with a SIMD score of 1 or 2 (most deprived). Additionally, post-op performance status (scale 0 - 5) accounts for two of the top ten feature values returned suggesting it is valuable predictor (see Table 7 in Appendix for post-op performance scale). Post-op performance status assumes the patient has had surgery and a status of 0 or 1 indicates a person is fully active or restricted only with strenuous physical activity following surgery. Comparatively, a status of 4 or 5 indicates a person is completely disabled or dead, respectively. In our dataset, 70% of patients with a post-op performance score of 0 or 1 survive more than a year compared to 37% of patients with a status of 2-5. The majority of the 10 features selected were used by at least one of the trained models.

Rank	Feature
1	Diagnosis: Glioma Benign
2	Diagnosis: Meningioma Benign
3	Diagnosis: Rare Tumour Benign
4	Age: (0.0, 44.0]
5	SIMD: 1.0
6	KP Score ≤ 70
7	Morphology: Heterogeneous
8	Post-op Performance Status: 0.0
9	Post-op Performance Status: 1.0
10	Urgency of Referral: Suspicion of Cancer (within 2 weeks)

Table 3: Top 10 features returned using sequential feature selector.

4.2.2. Feature Importance

An examination of a model's feature weightings can give rise to simple interpretations of how a classification is made. Although not fully transparent, this method allows moderate insight into how a model works and may assist clinicians in understanding causal factors for patient survival. The Cox PH model provides means for feature importance analysis. Notably, the Cox PH model is seldomly used to determine feature importance, rather it is used as a modelling technique to help adjust for confounders while assessing the effect side of a variable of interest. However, in the present case, the Cox PH model serves as an additional resource for feature comparison. Additionally, although SVM is the best performing model, the algorithm employs a kernel transformation hence feature importance cannot be directly computed. Thus the feature importance of RF, the second best model, was assessed using permutation importance [32].

The Cox PH model uses the variable's coefficient and standard error value to assess feature importance as illustrated in Figure 5. A feature value above 0 indicates increased risk, thus reduced survival time, whereas a value less than 0 suggests reduced risk, or increased survival time. A coefficient value of 0 indicates the feature has no importance. It is worth remarking, as discussed in Section 3.2, the first category per feature was dropped to account for problems with high collinearity. For example, the feature *Tumour Morphology* contains the values *Heterogenous* and *Homogenous - Heterogenous* was dropped. Thus Cox PH model feature importance should be interpreted with caution as not all feature types are present. However, general comparisons can be made.

According to the Cox PH model, a poor post-op status, no treatment and an older age at diagnosis are poor prognostic factors. Comparatively, chemotherapy, homogeneous tumour morphology and 100% tumour resection are positive prognostic factors. Similar features were used by both the rule lists and sequential feature selector. Interestingly, a malignant meningioma was a positive prognostic factor. However, the 1-year survival rate of malignant meningioma's is around 80% [82] thus highlighting the importance of the survival threshold

when assessing prognostic variables.

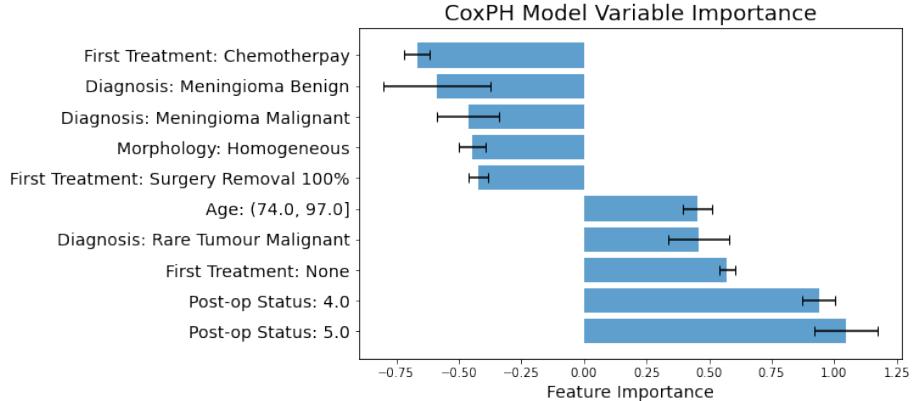


Figure 5: Bar chart showing the average Cox PH model feature importance. The coefficients were averaged across cross-validation folds and the standard deviation is shown with black error bars. The top 5 features with the largest negative coefficients and largest positive coefficients (i.e. the most informative features) were plotted.

RF feature importance is given in Figure 6. This was obtained using permutation importance, which measures the decrease in a model’s performance when a feature value is randomly shuffled [32] but does not indicate whether the feature is positively or negatively correlated with the predicted outcome. Given our dataset, which contain features with cardinalities ranging from 2 (e.g. Sex) to 9 (e.g. Diagnosis), permutation importance was preferred over the frequently-used mean decrease in impurity (MDI) [83] because the latter is often biased towards features with high cardinality [84]. As can be seen, the three most influential features for the RF model were a younger age at diagnosis, a post-op status of 0 and heterogeneous tumour morphology. All three features were also used by the rule lists and sequential feature selector. Note the three feature values were not used to train the Cox PH model (due to collinearity problems as discussed previously), however the Cox PH model found the inverse feature values (i.e. an older age at diagnosis, a post-op status of 4/5 and homogeneous tumour morphology) to be informative.

Notably, the Cox PH model is seldomly used to determine feature impor-

tance, rather it is a modelling technique to help adjust for confounders while assessing the effect side of a variable of interest.

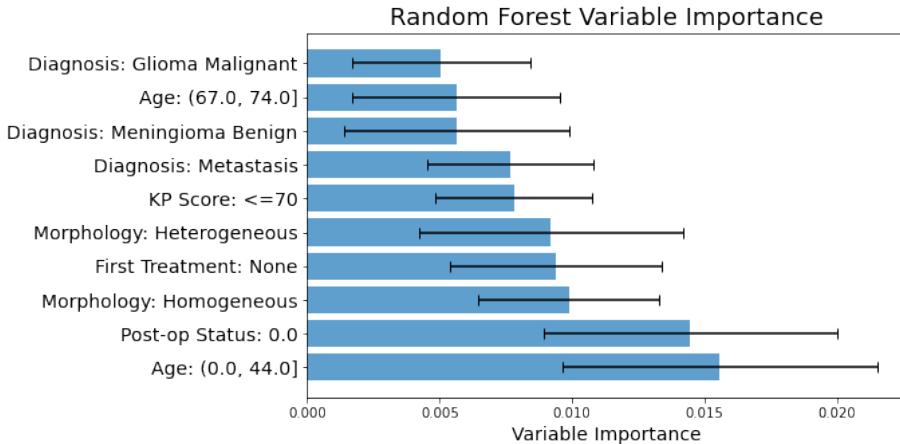


Figure 6: The 10 most influential features from the RF algorithm averaged across cross-validation folds. The standard deviation is shown with black error bars. Note that the importance does not specify positive or negative correlation.

4.2.3. Local Surrogate Model

Compared to feature importance, LIME and SHAP can assess interpretability at the local level for individual predictions, rather than at the global modular level. Local explanations may be more accurate than global explanations [26] and are beneficial for understanding why instances are classified incorrectly. For example, Figure 7 illustrates LIME’s explanation of an observation that was classified incorrectly by LR and SVM (i.e. wrongly predicted survival > 1 year) and correctly by RF (i.e. rightly predicted survival < 1 year). The top 10 influential features for the given test instance are shown. Figure 8 shows an explanation by SHAP of the same observation classified by RF. Both surrogate models were applied to all three ML models, and the same prediction instance was compared across the three models (see Figures 11 and 12 in the Appendix for LR and SVM explanation by SHAP). Note that the three ML models used one-hot encoded data to make predictions thus both the presence (value = 1)

and absence (value = 0) of a feature is used to make a prediction. Across all six model explanations (three ML models by two surrogate methods), similar features were used but the weighting of the features varied. For example, according to LIME, all three ML models placed the most importance on a younger age *not* being present, and for LR and SVM this was followed by *not* having a post-op status of 0 and *not* having a metastatic tumour. In comparison, RF valued the presence of a heterogeneous tumour, *not* a homogenous tumour and *not* having a benign meningioma. Additionally, according to LIME, six features were found to be the most influential across all models: age, morphology, diagnosis, post-op status, KP score and first treatment. In comparison, according to SHAP for the same test instance classified by RF (Figure 8), features such as lobe, midline shift, SIMD and symptom type were in the top 10 most influential features. The discrepancy between LIME and SHAP for the same model highlights the unreliability of post-hoc interpretability methods [22].

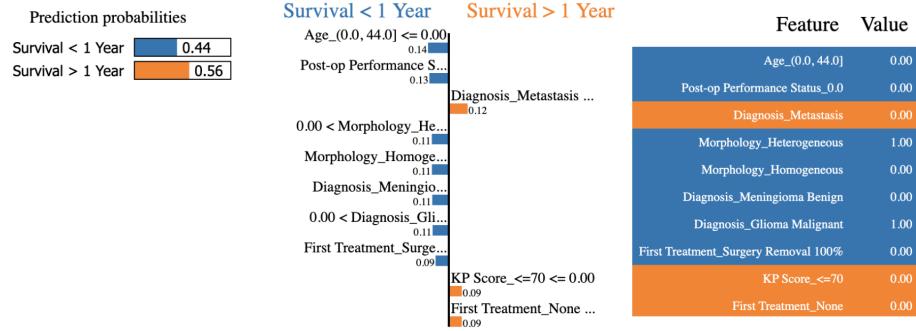
4.2.4. Qualitative Analysis

In medicine, where human decision making governs the process of patient treatment, a survey over domain experts is a valuable measure of interpretability [23]. The interpretability of our rule lists was assessed based on the clinical expertise of two of the current authors (M.P. and P.B.), who were provided with multiple point-estimates from both BRL and FRL models (see Figure 3, Figure 4 and Appendix B for all rule lists provided). To mitigate any potential bias, the models were constructed without their input and only the final models were presented for relatively informal feedback on their potential clinical utility as a predictive model.

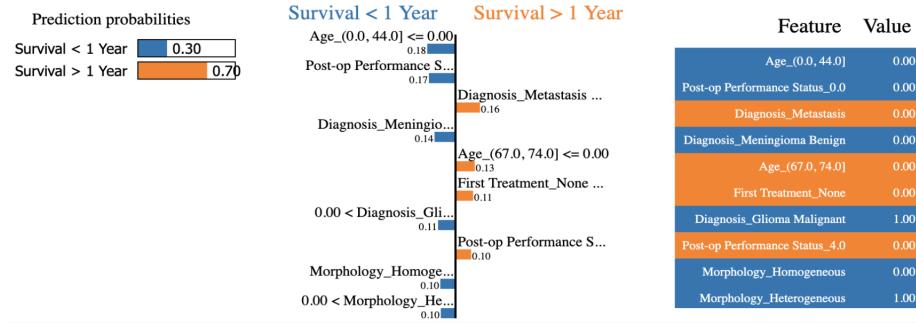
According to this evaluation, the combination of features used by the rule lists for survival prediction are informative and in-line with domain knowledge. Similar features were found influential across multiple rule lists including age at diagnosis, KP score, tumour morphology, and first cancer treatment. As expected, the diagnosis (tumour type) is also an important factor for survival. For example, benign gliomas and benign meningiomas indicated greater survival,



(a) Random Forest (RF) feature importance for a given test instance determined by LIME. The observation was correctly classified.



(b) Logistic Regression (LR) feature importance for a given test instance determined by LIME. The observation was incorrectly classified.



(c) Support Vector Machine (SVM) feature importance for a given test instance determined by LIME. The observation was incorrectly classified.

Figure 7: Negative (blue) features indicate survival less than a year, and positive (orange) features indicate survival greater than a year. The top 10 most influential features for the specific test instance are shown. The weight of each feature (centre image) is used to calculate the prediction probability.

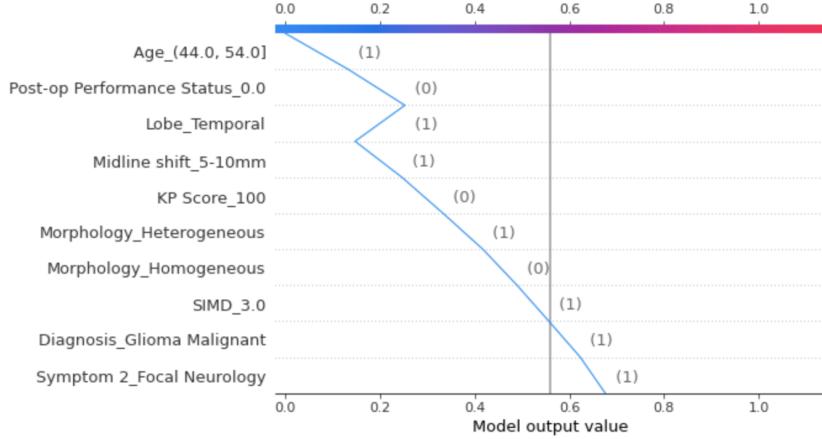


Figure 8: RF feature importance determined by SHAP. The top 10 most influential features for the specific test instance are shown. The prediction starts at a baseline value (0.410) which is the average of all predictions (not shown here). Moving from the bottom of the plot to the top, each Shapley value is added to the model’s base value and either pushes to increase (positive value) or decrease (negative value) the prediction.

both of which are highly treatable, with a 76% [2] and 70% [85] five-year survival rate, respectively. Comparatively, brain metastases indicated poor survival and only have a 1-year survival rate of 17% [86]. The impact of these feature combinations have been well-established in the literature [46, 47, 75, 76] and were re-produced by the rule-lists. An agreement between model predictions and clinical knowledge is essential for establishing trust in the models decision making process, as well as trust that the model will make accurate predictions when applied to new data.

4.3. Model Comparison

The features found important by the three model types (Cox PH, rule lists, ML) for survival prediction are summarised visually by a Venn diagram in Figure 9. For ease of evaluation, the general feature types are used for comparison between models (e.g. sex) instead of the individual feature values (e.g. female or male). Additionally, the features used by the ML models are for a single test instance thus only general comparisons can be made.

All of the features returned by the sequential feature selector (see Table 3), except *Urgency of Referral*, were used by at least one of the three model types. All features used by the baseline Cox PH model were also found important by the rule lists and ML models. We only investigated the top 10 most influential features used by the Cox PH model hence there is likely more overlap in feature significance between the three models. The features identified as influential by all models included age, diagnosis, morphology, first treatment and post-op status. The rule lists and ML models had multiple overlapping features including comorbidity, KP Score, symptom 1, symptom 2 and midline shift. Only the rule lists found sex and history of cancer to be significant, while only the ML models found SIMD, symptom 1 duration and lobe to be important. The importance of symptomatology data is interesting, as symptoms are often talked of in regards to time to diagnosis, but less often with regards to prognosis. When looking at the RF global top 10 important features (see Figure 6), symptom and sign features were not found relevant, but when using SHAP to look at a specific test instance from the RF model (see Figure 8), as well as LR and SVM (see Figure 11 and 12 in Appendix), symptom information was relevant. As mentioned above, sex was only found relevant for the rule lists which was surprising. Majority of research into sex differences and brain tumour survival have focused on glioblastomas, which find that men are more likely to develop, and die of, glioblastomas than women [87, 88, 89]. In comparison, women are twice as likely to develop meningiomas compared to men but no significant difference in outcome have been reported [89, 90]. Although sex may play an important role in individual tumour types, across all tumour types sex may be less influential. Finally, four features were not used by any model: Urgency of Referral, Sign 1, Side and Max Size. Note Urgency of Referral was found to be important by the sequential feature selector (see Table 3) but not by the algorithms. Given the importance of symptom 1 and symptom 2, the irrelevance of sign 1 (the first onset objective clinical abnormality) is perhaps the most surprising. Although in our dataset 42% of patients presented with no signs, as symptoms dominate, hence this feature may be less informative for classification.

Features which are not ranked as important by models should arguably be given less attention in clinical evaluation and the use of such features should be reviewed if they currently play a decision role in the assessment. It is beneficial to simplify models, and decision making, by reducing some of the noise over features, as there is often extraneous data in healthcare that can impair decision making rather than support it.

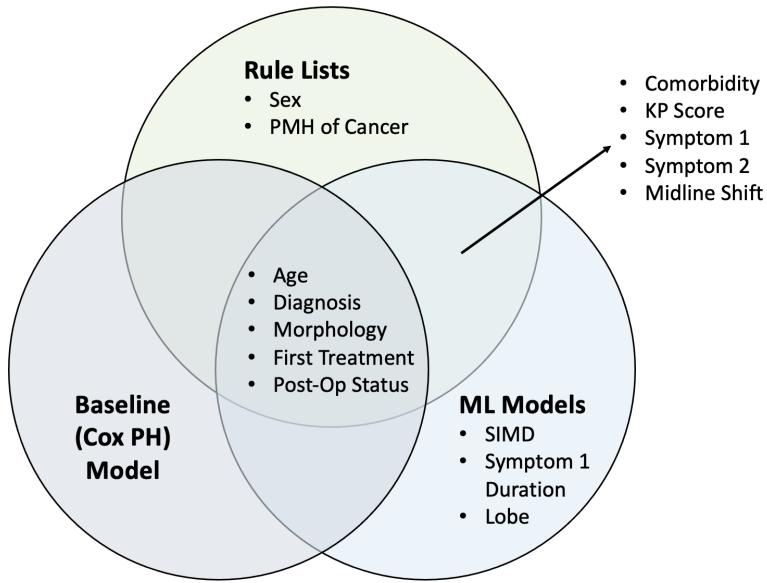


Figure 9: Venn diagram of the features used by the three main groups of models for survival prediction. Features used by all model types are found in the centre. Note the general feature types are compared between models instead of the individual feature values.

Model interpretability offers valuable insight on the importance of the individual features for survival prediction. Although there is an overlap in the features used by the models, the weighting of each feature differs and a comparison of this weighting may prove the most informative as it helps one understand what drives the model's performance. This is illustrated best with LIME and SHAP where for a given test instance the weightings of each feature used for classification are returned. For example, for the test instance discussed in Section 4.2.3, where RF correctly classified the instance but LR and SVM did not,

the weighting of the feature *not having a metastatic tumour* varied. According to LIME, RF placed a weight of 0.09 on not having a metastatic tumour, compared to LR and SVM which gave this feature a weighting of 0.12 and 0.16, respectively. Additionally, according to SHAP, this feature was *not* in the RF’s top 10 informative features, but it was for LR and SVM. Hence an understanding of a features predictive power and how this varries between models is valuable for prognostic decision making.

5. Discussion

The results summarized in Table 2 demonstrate that the BRL and FRL algorithms were outperformed slightly by popular ML models. However, the innate interpretability of rule lists may mitigate potential performance loss, especially in health-care where model transparency is essential for its integration into- and influence on decision making [22]. The original BRL paper derived a stroke prediction model and found that the BRL-point estimate performed on par with SVM and only $\sim 2\%$ worse than RF (their best performing model). Comparatively, the average BRL performance for brain tumour survival prediction was $\sim 5\%$ worse than SVM (our best performing model). Notably the dataset used by the original authors was more than $12\times$ the size of ours, hence the additional training data may have helped BRL performance. The minimal decrease in BRL performance for both stroke and brain tumour survival classification highlights the potential of innately interpretable algorithms for high-stake predictions.

Our results are also similar to that of Wang and Rudin [31], who found that FRL performance was slightly worse, but comparable to RF, LR and SVM on four public datasets. FRL performance on our dataset is within one standard deviation of the ML algorithms mean performance. The minimal loss in FRL performance is not surprising as the model’s strong monotonicity constraints may sacrifice performance [31].

More generally, the slight loss in rule list performance compared to the ML models may be due to our dense heterogeneous dataset and limited sample size.

Rule list algorithms optimise over pre-mined rules, rather than the entire feature space, hence the algorithm’s computational effort scales with the number of antecedents rather than the number of features [30]. Thus in general, less computation is required when the data is sparse. For BRL, depending on the hyperparameters, the number of antecedents used ranged from ~ 850 to ~ 2500 across the training folds (this information was not available for FRL). Although we attempted to reduce dataset complexity (e.g. by grouping rare tumours together and combining symptom and sign data into larger domains), rule list models may be better suited for a dataset with sparse features, especially given a small sample size. A model with higher statistical capability, such as the ML models or neural networks, may be better suited for handling dense features with limited training examples, however such models are often the least transparent.

Rule-lists are a promising interpretable model but are currently limited to categorical features and binary classification (see [91] for a recent proposal of a multi-class rule list algorithm). Similar to decision trees, rule lists cannot deal with linear relationships and due to their categorical input requirement, rules can only produce step-like predictions [26]. Additionally, a caveat to rule list interpretability is that such models are generally not stable. There may be multiple equally good rule lists and it is not clear which will be returned by the algorithm [30]. Furthermore, small alterations in the training dataset can result in a different rule list, which is especially problematic given our small dataset size. Thus using the full posterior of rule-list samples (i.e. BRL-post) rather than a single rule-list (i.e. BRL-point) may improve model stability, however, the classifier would no longer be interpretable. Despite these limitations, rule-lists are easy to interpret, and only the relevant features for rule list construction are selected. The classification of a patient is fast, whereby only a few binary statements need to be reviewed. BRLs offer a more fine-grained approach than FRLs, whereby rules that favour survival less than and greater than a year are used. In comparison, FRLs have the added benefit of automatically stratifying patients by risk thus a clinician only needs to look at the top few rules to

classify the most high risk patients. Similar to Senders et al. [16], an interactive graphical representation of the rule list model could be created, whereby a clinician inputs a patient’s features and a prediction is returned along with the decision rules used to make that prediction. Finally, the integration of additional clinical information such as blood tests or genetic data may improve the clinical validity of rule list models. Blood tests are now being investigated as a means for brain tumour diagnosis [92, 93] and genetic alterations have shown to be effective predictors for tumour prognosis [14, 94].

Although the innate interpretability of rule-lists cannot be directly compared to the post-hoc interpretability of ML algorithms, general conclusions can be made. Traditional variable importance algorithms are limited to group-level analysis but LIME and SHAP enable the explanation of individual predictions to understand the contribution of each predictor. Our results showed that similar features were used by the three ML models for making predictions. However, the weighting of each feature varied between models and across interpretability techniques. The different feature weightings between models may explain the variability in model performance, but the difference within interpretability techniques highlights the inconsistency between post-hoc methods. Because post-hoc methods assess interpretability after model construction, explanations can prove misleading and unreliable [22], and there is a growing body of literature that has questioned the credibility of LIME and SHAP [22, 95, 96, 97, 98]. There is a large toolbox of post-hoc methods [26] available to assess model interpretability, however, the absence of a cohesive definition for interpretability makes it difficult to assess the quality of different interpretability techniques. By agreeing on clear quantitative metrics for interpretability, the development of robust trusted interpretability techniques can follow [99].

5.1. Strengths and Limitations

This was the first study to apply classification algorithms to a novel brain tumour dataset for 1-year survival prediction. Brain tumours are a rare but deadly disease, and compounded by their heterogeneity, an accurate progn-

sis by clinicians is a challenging task. Also, to our knowledge, this is the first study to compare intrinsic and post-hoc interpretability methods for the assessment of predictive brain tumour survival models. As ML continues to advance, the development of tools to assist clinicians with these high-stake medical decisions by providing trustworthy data-driven support will become essential for acceptability. As we have already discussed, some of the advantages of rule lists in clinical practice are as follows: their innate interpretability, their simple *if...then...* structure is easy to follow and predictions with rules are fast (only a few binary statements need to be assessed).

Nonetheless, several important limitations remain. In the present study, only clinical prognostic factors were used for the prediction of survival. As previously mentioned, blood tests and molecular genetic alterations have been recognised as powerful prognostic and predictive markers in brain tumour survival [92, 93, 94, 100, 101, 102]. Hence the integration of the different data types may improve current survival predictions. However, an increase in data dimensionality is an important consideration especially when rule lists are employed. In addition, the data used in this study was heterogeneous and a second validation dataset is required for confirmation of our results. With the addition of new data, this study could also be extended to explore the application of post-hoc interpretability methods to neural networks. As discussed above, rule-lists currently require categorical features and are limited to binary classification (see [91] for multi-class rule list algorithm proposal). Extension of the rule list algorithm for multi-class classification or regression is an important next step for improving rule-list performance and constructing a competitive interpretable rule list classifier.

Finally, although not directly related to the ML models, our dataset required imputation and discretisation. There is the potential for imputation to introduce bias into the data [38, 103] and the chosen imputation method can influence the final results [104]. Additionally, although discretisation can be used to improve the clarity of classification models by extracting useful feature intervals, the split of the feature will also effect a model’s performance. Both pre-processing

techniques have the potential to reduce model performance which we sought to mitigate during dataset preprocessing through the exploration of multiple imputation and discretisation algorithms.

6. Conclusion

This study investigated the performance and interpretability of multiple algorithms for the prediction of brain tumour survival on a novel dataset. We have demonstrated that rule list algorithms create reliable and understandable results that have clinical relevance without significant compromise in model performance. Rule lists and other interpretable models may provide an advantage over traditional clinician assessment of prognosis by weighting potential risk factors and stratifying patients accordingly. Additionally, the slight superiority in ML model performance may be less important than the agreement of features between different models within a clinical context since the latter can provide more confidence as to the importance of those features. As shown in the current work, the reliability of LIME and SHAP for assessing feature importance is questionable and the methods are vulnerable to failures. Interpretability is crucial for the implementation of ML algorithms in healthcare, because prediction tools inform, but do not single-handedly direct, clinical decision making. A model’s ability to explain its predictions is essential for establishing a user’s trust in the model. The interpretable models introduced in this work attempt to bridge the gap between ML research and integration into clinical practice. Rule lists are not meant to be a direct competitor for black box classifiers, but rather a useful tool that can assist humans with high-stake decisions by providing trustworthy data-driven support. Further clinical utility in the current context may come from using other interpretable approaches with clinical and molecular (or imaging) data, where it is difficult for a clinician to determine what the most informative features are across the different data sources. Interpretable models are a natural choice for the domain of predictive medicine, and whether the model is innately interpretable or post-hoc methods are utilised, the validation

and integration of such models into clinical practice is an important next step for improving patient outcomes in a trusted way.

Funding

MTCP is supported by Cancer Research UK Brain Tumour Centre of Excellence Award (C157/A27589).

Conflict of interest statement

The authors declare no conflict of interest.

Appendix A

6.1. Glossary of Dataset Features

Table 4: Overview of dataset variables including their descriptions, value and percentage of each value present in the final dataset following imputation and discretisation.

Name	Description	Value	Proportion (%)
Age	the age of a patient	0-44	17.1
		45-54	18.7
		55-61	16.0
		62-67	15.8
		68-74	16.9
		75+	15.5
Sex	the sex of the patient	Male	50.5
		Female	49.5
History of Cancer	whether the patient has a past medical history of cancer	Yes	18.2
		No	81.8
Comorbidity	the presence of another illness or disease occurring in a patient	Yes	47.7
		No	52.3
Scottish Index of Multiple Deprivation (SIMD)	a measure of deprivation of the area a patient lives from most deprived (ranked 1) to least deprived (ranked 5)	1	13.9
		2	22.6
		3	21.0
		4	18.8
		5	23.7
Karnofsky Performance Score (KP Score)	a common measure in oncology to assess the functional state of a patient	100	37.6
		90	28.6
		80	14.6
		≤70	19.2

Continued on next page

Table 4 – continued from previous page

Name	Description	Value	Proportion (%)
Symptom 1	the first symptom type a patient presented with (reported by the patient)	Focal Neurology Headache Fits/Faints/Falls Behavioural/Cognitive Other/Non-specific Non-specific Neurological	34.6 28.4 17.1 16.7 2.4 0.8
Duration	Symptom 1 Duration (the length of time of a patient's first symptom)	0-2 weeks 3-4 weeks 5-8 weeks 9-20 weeks 20-52 weeks	20.6 20.1 19.5 20.4 19.4
Symptom 2	the second symptom type a patient presented with (reported by the patient)	Focal Neurology No Symptoms Behavioural/Cognitive Fits/Faints/Falls Headache Other/Non-specific	31.3 30.4 18.9 9.1 6.4 3.9
Sign 1	the first sign type a patient presented with (observed by the physician)	No Signs Neurological Cognitive Cranial Nerve Other Behavioural	42.7 36.2 15.0 5.0 0.8 0.3
Urgency of Referral	the patient's urgency of referral from primary care	Emergency Suspicion of Cancer (within 2 weeks) Soon (up to 3-4 weeks) Routine (up to 12 weeks)	59.7 17.3 2.9 20.1

Continued on next page

Table 4 – continued from previous page

Name	Description	Value	Proportion (%)
Diagnosis (or Tumour Type)	the type of brain tumour a patient was diagnosed with	Glioma Malignant	46.5
		Metastasis	19.0
		Meningioma Benign	13.6
		Glioma Benign	7.1
		Rare Tumour Benign	4.7
		Lymphoma Malignant	4.1
		Meningioma Malignant	2.3
		Rare Tumour Malignant	1.5
		Hemangioblastoma	1.2
		Benign	
Max Size	a measure of the tumour size	≤ 20	19.7
		21-40	38.1
		41-60	30.8
		≥ 61	11.4
Side	the side of the brain the tumour is located	Left	41.9
		Right	41.2
		Both Left and Right	11.6
		Midline	5.3
Lobe	the lobe where the tumour is located	Frontal	34.2
		Temporal	21.6
		Parietal	14.6
		Multiple	12.2
		Cerebellar	7.3
		Brainstem	5.7
		Occipital	4.4
Morphology	the histological classification of the tumour based on the cell types present	Heterogenous	68.5
		Homogenous	31.5

Continued on next page

Table 4 – continued from previous page

Name	Description	Value	Proportion (%)
Midline Shift	a measure of the tumour's horizontal shift from the mid (centre) line	0	43.3
		< 5mm	28.1
		5-10mm	17.4
		> 10mm	11.2
First Treatment	the type of first cancer treatment	Surgery Removal 100%	16.0
		Surgery Removal 90-99%	24.4
		Surgery Removal 50-89%	6.4
		Surgery Removal <50%	4.9
		Biopsy	16.9
		Radiotherapy	5.5
		Chemotherapy	0.9
Post-operative Performance Status	a measure of a patient's level of functioning following surgery in terms of their ability for self-care, daily activity, and physical ability	Other (e.g. steroids)	2.5
		No Treatment	22.5
		0	31.5
		1	27.4
		2	6.2
		3	1.9
Status	self-care, daily activity, and physical ability	4	1.4
		5	0.2
			No Surgery 31.4

6.2. Preprocessing

Condition	Percentage	Comments
A: Able to carry on normal activity and to work. No special care is needed.	100	Normal, no complaints, no evidence of disease.
	90	Able to carry on normal activity, minor signs or symptoms of disease.
	80	Normal activity with effort, some signs or symptoms of disease.
B: Unable to work. Able to live at home, care for most personal needs. A varying degree of assistance is needed.	70	Cares for self, unable to carry on normal activity or to do active work.
	60	Requires occasional assistance, but is able to care for most of his needs.
	50	Requires considerable assistance and frequent medical care.
C: Unable to care for self. Requires equivalent of institutional or hospital care. Disease may be progressing rapidly.	40	Disabled, requires special care and assistance.
	30	Severely disabled, hospitalization is indicated although death not imminent.
	20	Hospitalization necessary, very sick, active supportive treatment necessary.
	10	Moribund, fatal processes progressing rapidly.
	0	Dead.

Figure 10: The original description of the Karnofsky performance status given by Karnofsky and Burchenal [48].

Group	Symptom Domain	Symptom Examples
1	Headache	Headache
2	Behavioral/Cognitive	Confusion, memory loss, strange behaviour
3	Focal Neurology	Ataxia, vertigo, vision problems,
4	Fits, faints or falls	Seizure, collapse, convulsion
5	Non-specific neurological	Poor balance, dizziness, gait abnormality
6	Other/non-specific	Vomiting, lethargy, sweating

Table 5: Symptom domain classifications based on Ozawa et al. [53], with examples of symptom types in the brain tumour dataset.

Group	Sign Domain	Sign Examples
1	No signs	No signs
2	Behavioral	Behaviour signs anxiety (e.g. fast speech, tremor, voices anxiety, crying) Behaviour signs depression (e.g. voices low mood, crying) Behaviour (withdrawn/apathetic) - not depressed Behaviour (aggressive/paranoid) - not anxious
3	Cognitive	Cognitive - problems performing tasks (e.g. calculation, planning, VF) Cognitive - problems with memory (forgetfulness) Cognitive - reduced conscious level/drowsiness (reduced GCS) Cognitive - other non-specific confusion
4	Neurological	Dysphasia - Receptive Dysphasia - Expressive Dysarthria - slurred or slow or staccato Unilateral weakness (UMN type ≥ 2 of arm/leg/face) Unilateral numbness (≥ 2 of arm/leg/face, or spinothalamic type) Problems with dexterity/fine manipulation Problems walking/unsteadiness (weakness/numbness) Problems walking/ataxia Problems with visual acuity (unilateral or bilateral) Problems with visual field (unilateral or bilateral)
5	Cranial Nerve	Papilloedema Diplopia CN problems 3, 4 or 6 Nystagmus (unilateral or bilateral) Facial numbness/tongue numbness (CN 5) Facial weakness (CN 7) Reduced smell/taste (CN 1 or 7) Deafness (unilateral/bilateral) (CN 8) Problems swallowing (dysphagia) (CN 9, 10) Problems with volume of speech (dysphonia) (CN 10)
6	Other	Other

Table 6: Sign domain classifications based on clinical expertise of some of the current authors.
 All examples are from the Brain Tumour dataset.

Grade	Description
0	Fully active, able to carry on all pre-disease performance without restriction.
1	Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature, e.g., light house work, office work.
2	Ambulatory and capable of all self-care but unable to carry out any work activities; up and about more than 50% of waking hours.
3	Capable of only limited self-care; confined to bed or chair more than 50% of waking hours.
4	Completely disabled; cannot carry on any self-care; totally confined to bed or chair.
5	Dead.

Table 7: Description of a patient's performance status (or functional state) developed by the Eastern Cooperative Oncology Group [105].

6.3. Local Surrogate Model

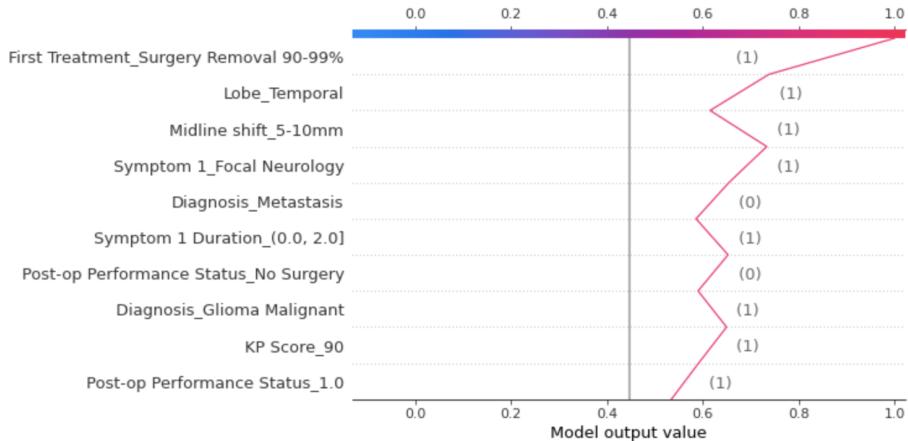


Figure 11: LR feature importance determined by SHAP.

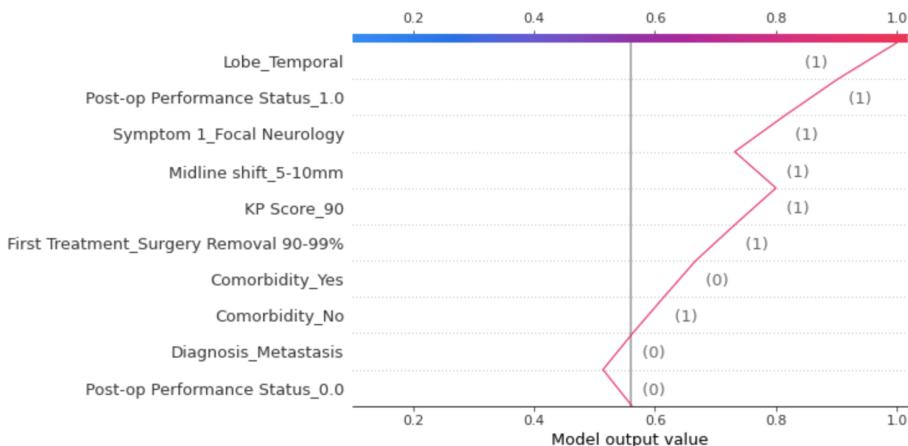


Figure 12: SVM feature importance determined by SHAP.

Appendix B

```
IF Symptom 1 : Behavioral/Cognitive AND KP Score : <=70 THEN probability of Survival >
1 year: 12.5% (5.6%-21.6%)
ELSE IF Diagnosis : Glioma Benign THEN probability of Survival > 1 year: 98.3%
(93.7%-100.0%)
ELSE IF Diagnosis : Meningioma Benign THEN probability of Survival > 1 year: 97.3%
(93.6%-99.4%)
ELSE IF Age : (0.0, 44.0] THEN probability of Survival > 1 year: 88.6% (81.3%-94.3%)
ELSE IF Sex : Male AND Morphology : Homogeneous THEN probability of Survival > 1
year: 81.1% (67.2%-91.8%)
ELSE IF First Treatment : None THEN probability of Survival > 1 year: 9.0% (4.2%-15.3%)
ELSE IF Age : (67.0, 74.0] AND Comorbidity : Yes THEN probability of Survival > 1 year:
14.0% (5.4%-25.6%)
ELSE IF Symptom 2 : Behavioral/Cognitive THEN probability of Survival > 1 year: 46.6%
(35.3%-58.0%)
ELSE IF Post-op Status : 0.0 THEN probability of Survival > 1 year: 77.1% (68.2%-84.9%)
ELSE IF First Treatment : Surgery Removal 90-99% AND Diagnosis : Glioma Malignant
THEN probability of Survival > 1 year: 73.9% (54.6%-89.3%)
ELSE probability of Survival > 1 year: 35.0% (27.4%-42.9%)
```

Figure 13: BRL-point estimate.

```

IF Diagnosis : Metastasis AND Midline shift : 0 THEN probability of Survival > 1 year:  

24.7% (15.8%-34.8%)  

ELSE IF First Treatment : None AND Morphology : Heterogeneous THEN probability of  

Survival > 1 year: 7.1% (3.0%-13.0%)  

ELSE IF Diagnosis : Glioma Benign THEN probability of Survival > 1 year: 98.3%  

(93.8%-100.0%)  

ELSE IF First Treatment : Surgery Removal 100% AND Sex : Female THEN probability of  

Survival > 1 year: 94.4% (88.1%-98.4%)  

ELSE IF Symptom 1 : Behavioral/Cognitive AND KP Score : <=70 THEN probability of  

Survival > 1 year: 15.9% (6.8%-27.9%)  

ELSE IF Diagnosis : Meningioma Benign THEN probability of Survival > 1 year: 97.3%  

(92.7%-99.7%)  

ELSE IF Age : (67.0, 74.0] AND Comorbidity : Yes THEN probability of Survival > 1 year:  

23.8% (12.4%-37.6%)  

ELSE IF Midline shift : 0 AND KP Score : 100 THEN probability of Survival > 1 year: 95.7%  

(89.8%-99.1%)  

ELSE IF Age : (0.0, 44.0] THEN probability of Survival > 1 year: 90.0% (79.1%-97.1%)  

ELSE IF Morphology : Homogeneous THEN probability of Survival > 1 year: 67.6%  

(51.3%-82.0%)  

ELSE IF KP Score : 90 AND Symptom 1 : Focal Neurology THEN probability of Survival > 1  

year: 75.6% (62.2%-86.8%)  

ELSE IF Post-op Status : 0.0 THEN probability of Survival > 1 year: 62.9% (50.6%-74.4%)  

ELSE probability of Survival > 1 year: 26.8% (19.4%-35.0%)

```

Figure 14: BRL-point estimate.

```

IF First Treatment : None AND Morphology : Heterogeneous THEN probability of Survival  

> 1 year: 8.1% (4.0%-13.5%)  

ELSE IF Diagnosis : Glioma Benign THEN probability of Survival > 1 year: 98.5%  

(94.4%-100.0%)  

ELSE IF Diagnosis : Metastasis THEN probability of Survival > 1 year: 39.5% (30.9%-48.4%)  

ELSE IF First Treatment : Surgery Removal 100% THEN probability of Survival > 1 year:  

94.8% (89.7%-98.3%)  

ELSE IF Symptom 1 : Behavioral/Cognitive AND KP Score : <=70 THEN probability of  

Survival > 1 year: 14.6% (5.7%-26.8%)  

ELSE IF Diagnosis : Meningioma Benign THEN probability of Survival > 1 year: 96.2%  

(89.6%-99.5%)  

ELSE IF Midline shift : 0 AND KP Score : 100 THEN probability of Survival > 1 year: 90.2%  

(81.6%-96.2%)  

ELSE IF PMH of Cancer : Yes THEN probability of Survival > 1 year: 31.6% (13.3%-53.5%)  

ELSE IF First Treatment : Surgery Removal 90-99% THEN probability of Survival > 1 year:  

76.1% (66.8%-84.4%)  

ELSE IF Sex : Female AND Comorbidity : Yes THEN probability of Survival > 1 year: 11.4%  

(3.3%-23.7%)  

ELSE probability of Survival > 1 year: 51.1% (42.8%-59.4%)

```

Figure 15: BRL-point estimate.

```

IF Diagnosis : Glioma Benign THEN probability of Survival > 1 year: 98.2% (93.6%-100.0%)  

ELSE IF Diagnosis : Meningioma Benign THEN probability of Survival > 1 year: 94.0%  

(89.0%-97.5%)  

ELSE IF Age : (0.0, 44.0] THEN probability of Survival > 1 year: 89.8% (82.7%-95.2%)  

ELSE IF First Treatment : None AND Morphology : Heterogeneous THEN probability of  

Survival > 1 year: 5.2% (2.0%-9.9%)  

ELSE IF Post-op Status : 0.0 AND Comorbidity : No THEN probability of Survival > 1 year:  

76.7% (67.3%-85.0%)  

ELSE IF Age : (67.0, 74.0] THEN probability of Survival > 1 year: 22.1% (14.0%-31.4%)  

ELSE IF Diagnosis : Metastasis THEN probability of Survival > 1 year: 37.9% (26.7%-49.8%)  

ELSE IF First Treatment : Surgery Removal 100% THEN probability of Survival > 1 year:  

95.0% (82.4%-99.9%)  

ELSE IF Midline shift : 0 AND KP Score : 100 THEN probability of Survival > 1 year: 93.1%  

(81.7%-99.1%)  

ELSE IF KP Score : 90 AND Sex : Male THEN probability of Survival > 1 year: 74.2%  

(57.7%-87.7%)  

ELSE probability of Survival > 1 year: 29.4% (22.2%-37.1%)

```

Figure 16: BRL-point estimate.

```

IF Morphology: Homogeneous AND KP Score: 100 THEN probability of Survival > 1 year is 98.54%, Support: 137
ELSE IF Age: (0.0, 44.0] AND PMH of Cancer: No THEN probability of Survival > 1 year is 92.86%, Support: 84
ELSE IF Morphology: Homogeneous AND PMH of Cancer: No THEN probability of Survival > 1 year is 77.38%, Support: 84
ELSE IF Post-op Status: 0.0 THEN probability of Survival > 1 year is 66.13%, Support: 124
ELSE probability of Survival > 1 year is 25.97%, Support: 385

```

Figure 17: FRL-point estimate.

```

IF Morphology: Homogeneous AND KP Score: 100 THEN probability of Survival > 1 year is 97.9%, Support: 143
ELSE IF Diagnosis: Meningioma Benign THEN probability of Survival > 1 year is 91.67%, Support: 48
ELSE IF Age: (0.0, 44.0] AND PMH of Cancer: No THEN probability of Survival > 1 year is 90.0%, Support: 70
ELSE IF First Treatment: Surgery Removal 100% AND PMH of Cancer: No THEN probability of Survival > 1 year is 84.62%, Support: 26
ELSE IF Post-op Status: 0.0 THEN probability of Survival > 1 year is 70.4%, Support: 125
ELSE probability of Survival > 1 year is 25.62%, Support: 402

```

Figure 18: FRL-point estimate.

```

IF Morphology: Homogeneous AND KP Score: 100 THEN probability of Survival > 1 year is 97.26%, Support: 146
ELSE IF Age: (0.0, 44.0] AND PMH of Cancer: No THEN probability of Survival > 1 year is 92.77%, Support: 83
ELSE IF Diagnosis: Meningioma Benign THEN probability of Survival > 1 year is 83.33%, Support: 36
ELSE IF First Treatment: Surgery Removal 100% AND PMH of Cancer: No THEN probability of Survival > 1 year is 83.33%, Support: 24
ELSE IF Post-op Status: 0.0 THEN probability of Survival > 1 year is 68.64%, Support: 118
ELSE probability of Survival > 1 year is 26.96%, Support: 408

```

Figure 19: FRL-point estimate.

```
IF Morphology: Homogeneous AND KP Score: 100 THEN probability of Survival > 1 year is  
97.26%, Support: 146  
ELSE IF Age: (0.0, 44.0] AND Comorbidity: No THEN probability of Survival > 1 year is  
93.44%, Support: 61  
ELSE IF Diagnosis: Meningioma Benign THEN probability of Survival > 1 year is 88.64%,  
Support: 44  
ELSE IF Post-op Status: 0.0 THEN probability of Survival > 1 year is 70.0%, Support: 130  
ELSE probability of Survival > 1 year is 30.25%, Support: 433
```

Figure 20: FRL-point estimate.

References

- [1] M. T. Poon, C. L. Sudlow, J. D. Figueroa, P. M. Brennan, Longer-term (≤ 2 years) survival in patients with glioblastoma in population-based studies pre-and post-2005: a systematic review and meta-analysis, *Scientific reports* 10 (2020) 1–10.
- [2] E. B. Claus, K. M. Walsh, J. K. Wiencke, A. M. Molinaro, J. L. Wiemels, J. M. Schildkraut, M. L. Bondy, M. Berger, R. Jenkins, M. Wrensch, Survival and low-grade glioma: the emergence of genetic information, *Neurosurgical Focus* 38 (2015) E6.
- [3] D. R. Cox, D. Oakes, *Analysis of survival data*, volume 21, CRC Press, 1984.
- [4] J. S. Barnholtz-Sloan, C. Yu, A. E. Sloan, J. Vengoechea, M. Wang, J. J. Dignam, M. A. Vogelbaum, P. W. Sperduto, M. P. Mehta, M. Machtay, et al., A nomogram for individualized estimation of survival among patients with brain metastasis, *Neuro-oncology* 14 (2012) 910–918.
- [5] H. Gittleman, D. Lim, M. W. Kattan, A. Chakravarti, M. R. Gilbert, A. B. Lassman, S. S. Lo, M. Machtay, A. E. Sloan, E. P. Sulman, et al., An independently validated nomogram for individualized estimation of survival among patients with newly diagnosed glioblastoma: NRG oncology RTOG 0525 and 0825, *Neuro-Oncology* 19 (2017) 669–677.
- [6] H. Gittleman, A. E. Sloan, J. S. Barnholtz-Sloan, An independently validated survival nomogram for lower-grade glioma, *Neuro-oncology* 22 (2020) 665–674.
- [7] T. Gorlia, M. J. van den Bent, M. E. Hegi, R. O. Mirimanoff, M. Weller, J. G. Cairncross, E. Eisenhauer, K. Belanger, A. A. Brandes, A. Allgeier, et al., Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE. 3, *The lancet oncology* 9 (2008) 29–38.

- [8] A. Iasonos, D. Schrag, G. V. Raj, K. S. Panageas, How to build and interpret a nomogram for cancer prognosis, *Journal of clinical oncology* 26 (2008) 1364–1370.
- [9] D. Bzdok, N. Altman, M. Krzywinski, Points of significance: statistics versus machine learning, 2018.
- [10] S. Bakas, M. Reyes, A. Jakab, S. Bauer, M. Rempfler, A. Crimi, R. T. Shinohara, C. Berger, S. M. Ha, M. Rozycki, et al., Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the BRATS challenge, arXiv preprint arXiv:1811.02629 (2018).
- [11] R. Jain, L. M. Poisson, D. Gutman, L. Scarpace, S. N. Hwang, C. A. Holder, M. Wintermark, A. Rao, R. R. Colen, J. Kirby, et al., Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor, *Radiology* 272 (2014) 484–493.
- [12] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, D. I. Fotiadis, Machine learning applications in cancer prognosis and prediction, *Computational and structural biotechnology journal* 13 (2015) 8–17.
- [13] P. Wang, Y. Li, C. K. Reddy, Machine learning for survival analysis: A survey, *ACM Computing Surveys* 51 (2019).
- [14] P. Fulop, A. Manataki, A. Agachi, P. Pop, Predicting survival after surgery for brain tumour patients: A machine learning study on clinical data and molecular data, In Proceedings of the AI for Social Good workshop, 7th International Conference on Learning Representations (ICLR 2019) (2019).
- [15] M. T. Ribeiro, S. Singh, C. Guestrin, “Why should i trust you?” Explaining the predictions of any classifier, *Proceedings of the ACM SIGKDD*

International Conference on Knowledge Discovery and Data Mining (2016) 1135–1144.

- [16] J. T. Senders, P. Staples, A. Mehrtash, D. J. Cote, M. J. Taphoorn, D. A. Reardon, W. B. Gormley, T. R. Smith, M. L. Broekman, O. Arnaout, An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning, *Neurosurgery* 86 (2020) E184–E192.
- [17] L. J. Wei, The accelerated failure time model: A useful alternative to the cox regression model in survival analysis, *Statistics in Medicine* 11 (1992) 1871–1879.
- [18] P. I. D’Urso, Letter: An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning, *Neurosurgery* 87 (2020) E273–E274.
- [19] X. Song, A. Mitnitski, J. Cox, K. Rockwood, Comparison of machine learning techniques with classical statistical models in predicting health outcomes., in: Medinfo, 2004, pp. 736–740.
- [20] M. A. Ahmad, C. Eckert, A. Teredesai, G. McKelvey, Interpretable machine learning in healthcare, *IEEE Intelligent Informatics Bulletin* 19 (2018) 1–7.
- [21] A. Holzinger, C. Biemann, C. S. Pattichis, D. B. Kell, What do we need to build explainable AI systems for the medical domain?, arXiv preprint arXiv:1712.09923 (2017).
- [22] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence* 1 (2019) 206–215.
- [23] S. Rüping, Learning interpretable models. Ph.D. thesis, Dortmund University of Technology (2006).

- [24] T. Miller, Explanation in Artificial Intelligence: Insights from the social sciences, *Artificial Intelligence* 267 (2019) 1–38.
- [25] D. V. Carvalho, E. M. Pereira, J. S. Cardoso, Machine learning interpretability: A survey on methods and metrics, *Electronics* 8 (2019) 832.
- [26] C. Molnar, Interpretable Machine Learning, 2019. <https://christophm.github.io/interpretable-ml-book/>.
- [27] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, N. Elhadad, Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission, in: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA, 2015, p. 1721–1730.
- [28] N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, D. Sontag, Population-level prediction of type 2 diabetes from claims data and analysis of risk factors, *Big Data* 3 (2015) 277–287.
- [29] C. Rudin, B. Ustun, Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice, *Interfaces* 48 (2018) 449–466.
- [30] B. Letham, C. Rudin, T. H. McCormick, D. Madigan, Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model, *Annals of Applied Statistics* 9 (2015) 1350–1371.
- [31] F. Wang, C. Rudin, Falling rule lists, *Journal of Machine Learning Research* 38 (2015) 1013–1022.
- [32] L. Breiman, Random forests, *Machine learning* 45 (2001) 5–32.
- [33] S. Menard, Applied logistic regression analysis, volume 106, Sage, 2002.
- [34] J. A. Suykens, J. Vandewalle, Least squares support vector machine classifiers, *Neural processing letters* 9 (1999) 293–300.

- [35] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: Advances in neural information processing systems, 2017, pp. 4765–4774.
- [36] R. Stupp, W. P. Mason, M. J. Van Den Bent, M. Weller, B. Fisher, M. J. Taphoorn, K. Belanger, A. A. Brandes, C. Marosi, U. Bogdahn, et al., Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma, *New England Journal of Medicine* 352 (2005) 987–996.
- [37] J. P. Klein, M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*, Springer Science & Business Media, 2006.
- [38] J. A. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, J. R. Carpenter, Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, *Bmj* 338 (2009).
- [39] L. Beretta, A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation, *BMC medical informatics and decision making* 16 (2016) 74.
- [40] P. D. Allison, Multiple imputation for missing data: A cautionary tale, *Sociological methods & research* 28 (2000) 301–309.
- [41] C. B. O’May, Investigating brain cancer survival with machine learning, University of Edinburgh (2019). Accessed from: https://project-archive.inf.ed.ac.uk/msc/20193500/msc_proj.pdf.
- [42] C. Zhang, S. Zhang, *Association rule mining: models and algorithms*, volume 2307, Springer, 2003.
- [43] E. A. Gehan, M. D. Walker, Prognostic factors for patients with brain tumors, *Natl Cancer Inst Monogr* 46 (1977) 189–195.
- [44] P. McKinney, Brain tumours: incidence, survival, and aetiology, *Journal of Neurology, Neurosurgery & Psychiatry* 75 (2004) ii12–ii17.

- [45] M. S. Walid, Prognostic factors for long-term survival after glioblastoma, *The Permanente Journal* 12 (2008) 45.
- [46] H. Gittleman, A. Boscia, Q. T. Ostrom, G. Truitt, Y. Fritz, C. Kruchko, J. S. Barnholtz-Sloan, Survivorship in adults with malignant brain and other central nervous system tumor from 2000–2014, *Neuro-oncology* 20 (2018) vii6–vii16.
- [47] S. Lapointe, A. Perry, N. A. Butowski, Primary brain tumours in adults, *The Lancet* 392 (2018) 432–446.
- [48] D. A. Karnofsky, W. H. Abelmann, L. F. Craver, J. H. Burchenal, The use of the nitrogen mustards in the palliative treatment of carcinoma. with particular reference to bronchogenic carcinoma, *Cancer* 1 (1948) 634–656.
- [49] D. Frappaz, A. Bonneville-Levard, D. Ricard, S. Carrie, C. Schiffler, K. H. Xuan, M. Weller, Assessment of karnofsky (kps) and who (who-ps) performance scores in brain tumour patients: The role of clinician bias, *Supportive Care in Cancer* (2020) 1–9.
- [50] A. Taylor, I. N. Olver, T. Sivanthan, M. Chi, C. Purnell, Observer error in grading performance status in cancer patients, *Supportive Care in cancer* 7 (1999) 332–335.
- [51] J. Sørensen, M. Klee, T. Palshof, H. Hansen, Performance status assessment in cancer patients. an inter-observer variability study, *British journal of cancer* 67 (1993) 773–775.
- [52] K. Chaichana, S. Parker, A. Olivi, A. Quiñones-Hinojosa, A proposed classification system that projects outcomes based on preoperative variables for adult patients with glioblastoma multiforme, *Journal of neurosurgery* 112 (2010) 997–1004.
- [53] M. Ozawa, P. M. Brennan, K. Zienius, K. M. Kurian, W. Hollingworth, D. Weller, R. Grant, W. Hamilton, Y. Ben-Shlomo, The usefulness of

symptoms alone or combined for general practitioners in considering the diagnosis of a brain tumour: a case-control study using the clinical practice research database (CPRD) (2000-2014), *BMJ Open* 9 (2019).

- [54] The Brain Tumour Charity, Adult brain tumour types (2020). <Https://www.thebraintumourcharity.org/brain-tumour-diagnosis-treatment/types-of-brain-tumour-adult> (Accessed: 1 August, 2020).
- [55] I. Dagogo-Jack, A. T. Shaw, Tumour heterogeneity and resistance to cancer therapies, *Nature reviews Clinical oncology* 15 (2018) 81.
- [56] A. Marusyk, K. Polyak, Tumor heterogeneity: causes and consequences, *Biochimica et Biophysica Acta (BBA)-Reviews on Cancer* 1805 (2010) 105–117.
- [57] S. Varma, R. Simon, Bias in error estimation when using cross-validation for model selection, *BMC bioinformatics* 7 (2006) 91.
- [58] M. Bekkar, H. K. Djemaa, T. A. Alitouche, Evaluation measures for models assessment over imbalanced data sets, *J Inf Eng Appl* 3 (2013).
- [59] M. Hossin, M. Sulaiman, A review on evaluation metrics for data classification evaluations, *International Journal of Data Mining & Knowledge Management Process* 5 (2015) 1.
- [60] A. P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern recognition* 30 (1997) 1145–1159.
- [61] C. Davidson-Pilon, J. Kalderstam, N. Jacobson, sean reed, B. Kuhn, P. Zivich, M. Williamson, AbdealiJK, D. Datta, A. Fiore-Gartland, A. Parij, D. WIlson, Gabriel, L. Moneda, A. Moncada-Torres, K. Stark, H. Gadgil, Jona, K. Singaravelan, L. Besson, M. S. Peña, S. Anton, A. Klintberg, GrowthJeff, J. Noorbakhsh, M. Begun, R. Kumar, S. Hussey, D. Golland, jlim13, Camdavidsonpilon/lifelines: v0.25.5, 2020. URL: <https://doi.org/10.5281/zenodo.4050560>.

- [62] G. Ambler, S. Seaman, R. Omar, An evaluation of penalised survival methods for developing prognostic models with rare events, *Statistics in medicine* 31 (2012) 1150–1161.
- [63] H. Liu, A. Gegov, M. Cocea, Rule based systems for big data: a machine learning approach, volume 13, Springer, 2015.
- [64] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the 20th International Conference on Very Large Data Bases, VLDB '94, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994, p. 487–499.
- [65] C. Borgelt, Frequent item set mining, *Wiley interdisciplinary reviews: data mining and knowledge discovery* 2 (2012) 437–456.
- [66] N. Hussein, A. Alashqur, B. Sowan, Using the interestingness measure lift to generate association rules, *Journal of Advanced Computer Science & Technology* 4 (2015) 156.
- [67] C. Borgelt, An implementation of the FP-growth algorithm, in: Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations, 2005, pp. 1–5.
- [68] W. K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Oxford University Press, 1970.
- [69] G. O. Roberts, A. F. Smith, Simple conditions for the convergence of the Gibbs sampler and Metropolis-Hastings algorithms, *Stochastic processes and their applications* 49 (1994) 207–216.
- [70] P. J. Van Laarhoven, E. H. Aarts, Simulated annealing, in: *Simulated annealing: Theory and applications*, Springer, 1987, pp. 7–15.
- [71] J. R. Quinlan, Induction of decision trees, *Machine learning* 1 (1986) 81–106.

- [72] A. Y. Ng, Feature selection, L1 vs. L2 regularization, and rotational invariance, in: Proceedings of the twenty-first international conference on Machine learning, 2004.
- [73] J. Park, I. W. Sandberg, Universal approximation using radial-basis-function networks, *Neural computation* 3 (1991) 246–257.
- [74] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, *the Journal of machine Learning research* 12 (2011) 2825–2830.
- [75] S. R. Dehcordi, D. De Paulis, S. Marzi, A. Ricci, A. Cimini, M. Cifone, R. Galzio, Survival prognostic factors in patients with glioblastoma: our experience, *J Neurosurg Sci* 56 (2012) 239–245.
- [76] F. G. Davis, B. J. McCarthy, S. Freels, V. Kupelian, M. L. Bondy, The conditional probability of survival of patients with primary malignant brain tumors: surveillance, epidemiology, and end results (SEER) data, *Cancer: Interdisciplinary International Journal of the American Cancer Society* 85 (1999) 485–491.
- [77] A. Stark, C. Stöhring, J. Hedderich, J. Held-Feindt, H. Mehdorn, Surgical treatment for brain metastases: Prognostic factors and survival in 309 patients with regard to patient age, *Journal of Clinical Neuroscience* 18 (2011) 34–38.
- [78] L. Woods, B. Rachet, M. Coleman, Origins of socio-economic inequalities in cancer survival: a review, *Annals of oncology* 17 (2006) 5–19.
- [79] B. Rachet, L. Ellis, C. Maringe, T. Chu, U. Nur, M. Quaresma, A. Shah, S. Walters, L. Woods, D. Forman, et al., Socioeconomic inequalities in cancer survival in england after the nhs cancer plan, *British journal of cancer* 103 (2010) 446–453.

- [80] M. Proctor, D. Morrison, D. Talwar, S. Balmer, D. O'reilly, A. Foulis, P. Horgan, D. McMillan, An inflammation-based prognostic score (mgps) predicts cancer survival independent of tumour site: a glasgow inflammation outcome study, *British journal of cancer* 104 (2011) 726–734.
- [81] A. Smith, D. Painter, D. Howell, P. Evans, G. Smith, R. Patmore, A. Jack, E. Roman, Determinants of survival in patients with chronic myeloid leukaemia treated in the new era of oral therapy: findings from a uk population-based patient cohort, *BMJ open* 4 (2014).
- [82] G. Truitt, H. Gittleman, R. Leece, Q. T. Ostrom, C. Kruchko, T. S. Armstrong, M. R. Gilbert, J. S. Barnholtz-Sloan, Partnership for defining the impact of 12 selected rare CNS tumors: a report from the cbtrus and the nci-connect, *Journal of neuro-oncology* 144 (2019) 53–63.
- [83] L. Breiman, Some properties of splitting criteria, *Machine Learning* 24 (1996) 41–47.
- [84] C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn, Bias in random forest variable importance measures: Illustrations, sources and a solution, *BMC bioinformatics* 8 (2007) 25.
- [85] B. J. McCarthy, F. G. Davis, S. Freels, T. S. Surawicz, D. M. Damek, J. Grutsch, H. R. Menck, E. R. Laws, Factors associated with survival in patients with meningioma, *Journal of neurosurgery* 88 (1998) 831–839.
- [86] K. Ekici, O. Temelli, M. Dikilitas, I. H. Dursun, N. Bozdag, E. K. Kaplan, Survival and prognostic factors in patients with brain metastasis: Single center experience, *Age* 230 (2016) 73.
- [87] M. Tian, W. Ma, Y. Chen, Y. Yu, D. Zhu, J. Shi, Y. Zhang, Impact of gender on the survival of patients with glioblastoma, *Bioscience reports* 38 (2018).

- [88] Q. T. Ostrom, J. B. Rubin, J. D. Lathia, M. E. Berens, J. S. Barnholtz-Sloan, Females have the survival advantage in glioblastoma, *Neuro-oncology* 20 (2018) 576.
- [89] T. Sun, A. Plutynski, S. Ward, J. B. Rubin, An integrative view on sex differences in brain tumors, *Cellular and molecular life sciences* 72 (2015) 3323–3342.
- [90] B. Holleczek, D. Zampella, S. Urbschat, F. Sahm, A. von Deimling, J. Oertel, R. Ketter, Incidence, mortality and outcome of meningiomas: a population-based study from germany, *Cancer epidemiology* 62 (2019) 101562.
- [91] H. M. Proen  a, M. van Leeuwen, Interpretable multiclass classification by MDL-based rule lists, *Information Sciences* 512 (2020) 1372–1393.
- [92] E. Gray, H. J. Butler, R. Board, P. M. Brennan, A. J. Chalmers, T. Dawson, J. Goodden, W. Hamilton, M. G. Hegarty, A. James, et al., Health economic evaluation of a serum-based blood test for brain tumour diagnosis: exploration of two clinical scenarios, *BMJ open* 8 (2018).
- [93] S. Podnar, M. Kukar, G. Gun  ar, M. Notar, N. Go  njak, M. Notar, Diagnosing brain tumours by routine blood tests using machine learning, *Scientific reports* 9 (2019) 1–7.
- [94] A. M. Molinaro, J. W. Taylor, J. K. Wiencke, M. R. Wrensch, Genetic and molecular epidemiology of adult diffuse glioma, *Nature Reviews Neurology* 15 (2019) 405–417.
- [95] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, M. Detyniecki, The dangers of post-hoc interpretability: Unjustified counterfactual explanations, arXiv preprint:1907.09294 (2019).
- [96] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods, in: Pro-

ceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186.

- [97] B. Dimanov, U. Bhatt, M. Jamnik, A. Weller, You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods., in: SafeAI@ AAAI, 2020, pp. 63–73.
- [98] E. Lee, D. Braines, M. Stiffler, A. Hudler, D. Harborne, Developing the sensitivity of LIME for better machine learning explanation, in: Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications, volume 11006, International Society for Optics and Photonics, 2019.
- [99] R. ElShawi, Y. Sherif, M. Al-Mallah, S. Sakr, Interpretability in health-care: A comparative study of local machine learning interpretability techniques, *Computational Intelligence* (2020).
- [100] J. E. Eckel-Passow, D. H. Lachance, A. M. Molinaro, K. M. Walsh, P. A. Decker, H. Sicotte, M. Pekmezci, T. Rice, M. L. Kosel, I. V. Smirnov, et al., Glioma groups based on 1p/19q, IDH, and TERT promoter mutations in tumors, *New England Journal of Medicine* 372 (2015) 2499–2508.
- [101] C. Hartmann, B. Hentschel, W. Wick, D. Capper, J. Felsberg, M. Simon, M. Westphal, G. Schackert, R. Meyermann, T. Pietsch, et al., Patients with IDH1 wild type anaplastic astrocytomas exhibit worse prognosis than idh1-mutated glioblastomas, and IDH1 mutation status accounts for the unfavorable prognostic effect of higher age: implications for classification of gliomas, *Acta neuropathologica* 120 (2010) 707–718.
- [102] B. H. Diplas, X. He, J. A. Brosnan-Cashman, H. Liu, L. H. Chen, Z. Wang, C. J. Moure, P. J. Killela, D. B. Loriaux, E. S. Lipp, et al., The genomic landscape of TERT promoter wildtype-IDH wildtype glioblastoma, *Nature communications* 9 (2018) 1–11.

- [103] J. C. Jakobsen, C. Gluud, J. Wetterslev, P. Winkel, When and how should multiple imputation be used for handling missing data in randomised clinical trials—a practical guide with flowcharts, *BMC medical research methodology* 17 (2017) 1–10.
- [104] M. R. Stavseth, T. Clausen, J. Røislien, How handling missing data may impact conclusions: A comparison of six different imputation methods for categorical questionnaire data, *SAGE open medicine* 7 (2019) 2050312118822912.
- [105] M. M. Oken, R. H. Creech, D. C. Tormey, J. Horton, T. E. Davis, E. T. McFadden, P. P. Carbone, Toxicity and response criteria of the eastern cooperative oncology group, *American journal of clinical oncology* 5 (1982) 649–656.