Building an Interpretable Machine Learning Classifier for the Prediction of Brain Tumour Survival

Colleen Charlton

Master of Science
Cognitive Science
School of Informatics
University of Edinburgh
2020

Abstract

Prognosis of a brain tumour is often poor, and the prediction of survival is one of the most challenging tasks physicians face. An accurate prognosis is essential for patient support and has a decisive impact on potential treatment regimens. This project uses a hand-curated brain tumour dataset to compare interpretable rule list models to a number of popular machine learning approaches for survival prediction. The performance of the interpretable models will be evaluated against "black box" machine learning models and the cost of transparency will be assessed. We investigate two rule list algorithms, Bayesian rule lists and falling rule lists, and qualitatively assess their interpretabiliy and clinical utility with the help of domain experts. We show that interpretable rule list models are able to predict survival on par with opaque machine learning models and have the added benefit of interpretability.

Acknowledgements

I would like to thank Dr. Jacques Fleuriot for his continuous support, constructive feedback and infectious enthusiasm throughout this project. I would also like to thank Dr. Paul Brennan and Michael Poon, both of whom provided invaluable medical knowledge over the course of this project. Finally, I would like to thank my family for their endless support, encouragement and love.

Table of Contents

Chapter 1

Introduction

1.1 Brain Cancer

Every year, approximately 12,000 people are diagnosed with a primary brain tumour in the United Kingdom (UK), which is equivalent to nearly 33 people a day [1]. A primary brain tumour originates in the brain itself, and the chance of developing this type of tumour is less than 1% [2]. Often malignant (cancerous) and benign (non-cancerous) tumours present with the same general symptoms. Symptoms may vary depending on the type, location and size of the tumour, and compounded by the rarity of the condition, brain tumours are one of the most difficult cancers to diagnose. Although the exact cause of a primary brain tumour is unknown, certain factors such as age or medical history may increase a person's risk [2]. Despite advancements in treatment, prognosis of a brain tumour is poor, with only 11% of adults surviving five years after diagnosis [1].

The decision to further investigate for a potential brain tumour is difficult. By the time a patient is referred to a specialist, the tumour may already be in an advanced state thereby limiting treatment options. The first course of treatment if often surgery, where the aim is to remove as much abnormal tissue as possible. This may be followed by radiotherapy or chemotherapy to treat any abnormal cells left behind. However even following treatment, a patient's outlook varies greatly. Brain tumours kill more children and adults under 40 than any other cancer in the UK [1]. Although survival rates are slowly improving [3], overall survival remains poor compared to most other cancers. Moreover, due to the rarity of brain tumours and range of brain tumour types, survival rates are difficult to predict. By identifying factors that influence survival, clinicians can tailor treatment regimens in an attempt to optimise patient outcome.

Recently, with the increased availability of clinical data, data-driven approaches such as machine learning (ML), have increasingly been applied to the medical domain. Such models may be developed to assist clinicians in symptom assessment, referral decisions and treatment plans. However, the implementation of ML models for high-stake medical problems first requires a user's understanding and trust in the model.

1.2 Objectives

The creation of a tool to assist clinicians in brain tumour diagnostic and treatment decisions has the potential to improve the current rate of patient survival. A physician's judgement is the final say, but ML models may provide novel insight that augments clinical expertise or gives support to a physician's decision.

This project will explore the benefits of interpretable ML approach's applied to the medical domain. Specifically, rule lists will be investigated as a type of interpretable ML model (introduced in Section 2.2). The purpose of this project is twofold. First, existing ML techniques will be applied to a novel brain tumour dataset, with the goal of implementing an accurate and interpretable ML model to predict patient survival. Second, the results about survival will be compared against other "black box" models, and the cost of transparency will be assessed. We also briefly explored the use of rule lists to predict a glioblastoma diagnosis, the most lethal type of brain cancer [4].

This project uses a hand-curated dataset collected by Dr. Paul Brennan, a Senior Clinical Lecturer and Honorary Consultant Neurosurgeon at the University of Edinburgh. Dr. Brennan also served as a clinical expert and was regularly consulted for guidance throughout this project.

1.3 Thesis Structure

The remainder of this thesis is organised in the following way. Chapter 2 will provide an overview of interpretability in ML, and introduce rule lists and other popular ML algorithms. This is followed by a discussion on discretisation and an overview of previous work on brain tumour survival prediction. Chapter 3 will introduce the dataset, including details of preprocessing, data imputation and discretisation. Chapter 4 will outline the methods and evaluation criteria used to assess the models. Finally, Chapter 5 will discuss the results and Chapter 6 closes with a discussion of final conclusions and future work.

Chapter 2

Background

This chapter will first discuss the importance of interpretable ML in high-stake decision making. This is followed by an introduction to rule list models and other ML algorithms, an overview of discretisation techniques, and a survey of related work on brain tumour survival prediction.

2.1 Interpretable Machine Learning

In the past, ML was employed for low-stake applications, such as online advertising, with arguably no significant impact on human life. Recently, with advancements in ML, algorithms are being used in high-stake ethical decisions such as criminal justice [5], finance [6] and heath care [7]. Notably in the medical domain, models are being developed to predict the risk of an event such as hospital re-admission [8], stroke outcome [9], or the development of cancer [10]. However, the employability of such models in a clinical setting comes with many legal and ethical considerations. Specifically, the model's lack of interpretability in its decision making process is a major limitation as end users (e.g. clinicians) will not act on a model's output in blind-faith.

There is no all-purpose definition of interpretability, as this is a subjective concept that is often domain-specific [11, 12]. Depending on the audience and application of the algorithm, different types of explanations may be warranted. A popular definition of interpretability by Miller is "the degree to which a human can understand the cause of a decision" [13]. Thus a model may be considered more interpretable than another model, if the prior's decisions are easier to comprehend than the latter's [14].

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

[11]. Often complicated black box models produce highly accurate predictions at the expense of human understanding. The idea that interpretability must be sacrificed for accuracy is flawed, and interpretability may even be used as a tool to improve accuracy [14]. For example, interpretable models by design provide insight into variable relationships and how final predictions are made. Unusual patterns in the data can be recognised, addressed appropriately, and thereby improve the model's accuracy. Moreover, the use of models with interpretability constraints have already shown to perform on par with unconstrained models across several health care domains [15, 16, 17].

2.1.1 Types of Interpretability

Interpretability methods can roughly be divided into two main techniques: intrinsic and post-hoc interpretability. Intrinsic interpretability refers to a model that by design is innately interpretable. The model is restricted to a simple structure allowing end-users to understand feature relationships and how final results are generated. Commonly used interpretable models include rule lists, decision trees and regression algorithms. However, the level of interpretability in these models vary, with rule lists being the most interpretable. Post-hoc interpretability refers to explanation techniques used to extract information from a trained model. The latter method has the advantage of flexibility, as these techniques can be applied to any model type, thus allowing developers the freedom to choose the model they desire. Post-hoc interpretability methods include feature importance, partial dependence or the use of a surrogate model (e.g. a linear model). For example, LIME (Local Interpretable Model-Agnostic Explanations) is popular surrogate model which locally applies a linear model to a prediction to understand how the prediction changes with data-point alterations [18]. When assessing multiple models, it may be advantageous to use post-hoc explanations, as the same technique can be applied to all model types. Although, post-hoc interpretability methods have proven to be powerful tools, they spawn the risk of generating explanations predicated on artefacts learned by the model rather than true patterns from the data [19]. Conversely, a model that is interpretable by design may yield more faithful explanations based on the model's own computations. Considering the high-stakes application of a predictive risk model, an algorithm which is inherently interpretable may be trusted more in clinical practice. Thus the focus of this thesis is on intrinsically interpretable ML methods, specifically rule-based models, however feature importance and LIME were also briefly experimented with (see Section 5.2).

2.2 Rule Lists

Rule lists have shown great success for many decades as a type of intrinsically interpretable model. This model produces a series of *if-then* rules, also known as decision rules, which are used to generate predictions. If a rule (or set of rules) is satisfied, the model outputs a certain classification. This general rule structure, *if* the conditions are satisfied *then* make a prediction, is semantically similar to natural language and thus appeals to human intuition. Although the concept of decision rules is well known, the use of ML to learn a rule list is relatively novel.

MYCIN, developed in the early 1970s, was an early rule-based expert system that used artificial intelligence (AI) to assist physicians in treatment decisions for infectious diseases [20]. MYCIN consisted of over 450 hand-curated rules that were developed using the heuristic knowledge of specialized domain-experts. This expert system would ask a series of questions designed to emulate the thinking of an expert, and also contained an explanation system that justified its recommendations. MYCIN was viewed as credible by professional users [21], and pioneered a new-stage of AI which utilised domain expertise and heuristic knowledge to solve a problem. The success of MYCIN spawned the development of many other medical expert systems [22, 23, 24].

Despite its success, MYCIN and other expert systems suffer from a number of limitations. The knowledge base provided by human experts may be incomplete, subject to bias and often lacks the common sense required for decision-making [25]. Expert systems are also expensive to build and maintain, and errors in the knowledge base may lead to wrong decisions. More recently, rule-based models are being constructed directly from datasets with the help of ML [26]. Instead of domain knowledge, the rules are *learned* straight from the data. This data-driven approach reduces the potential for human error, and is significantly more time and cost-effective.

There are many algorithms available to learn decision rules from data. This project will explore the Bayesian Rule List (BRL) [27] and Falling Rule List (FRL) [28] algorithms, described in Sections 2.3 and 2.4, respectively. Both BRL and FRL models are a type of generative algorithm, which explicitly models the distribution of each class, compared to a discriminative algorithm which models the decision boundary between classes. Other decision rule algorithms such as OneR [29] and Sequential Covering [30], will not be reviewed in this thesis and a discussion on the fundamentals of rule learning can be found elsewhere [31].

2.3 Bayesian Rule Lists

BRLs combine pre-mined frequent patterns from the dataset into a decision list using Bayesian statistics [27]. Bayes' theorem is a simple formula for calculating conditional probabilities. It can be used for the classification of data, referred to as a Naive Bayes Classifier [32], such that:

$$P(C|X) = \frac{P(X|C)P(C)}{P(X)} \tag{2.1}$$

where $P(C|X)$ is the posterior probability of a class given some data, $P(X|C)$ is the likelihood of the data given the class, $P(C)$ is the prior probability of the class, and $P(X)$ is the probability of the data. For classification, $P(X)$ is effectively constant and can be ignored. The model makes a 'naive' assumption of conditional independence where it is assumed that the feature probabilities are independent given the class. Given the training data and specified hyperparameters, BRL's use this theorem to create a probabilistic classifier that optimises the posterior over rule lists [27].

BRL's are used for classification problems where the goal is to learn $P(Y = 1 \mid X)$. Y is binary, and in the case of predicting a brain tumour, $Y = 1$ would indicate the presence of a brain tumour and X would represent a patient's features. The conditional probability distribution is represented as a decision list consisting of a series of *if...then...* rules. Figure 2.1 shows an example of a BRL constructed from our dataset which predicts the probability a patient will survive more than one year following a brain tumour diagnosis. The creation of a BRL roughly follows the subsequent steps: first, antecedents are extracted from the data using a rule mining technique and second, a set of rules and their order are learned using Bayesian statistics. The steps are discussed in detail below.

IF Diagnosis: Meningioma **AND** Treatment: Excision surgery **THEN** probability of Survival > 1 year : 95.7% (90.9%-98.8%)

ELSE IF Age: 19-50 **AND** Urgency of referral: Emergency **THEN** probability of Survival > 1 year : 96.2% (89.7%-99.5%)

ELSE IF Resection: No Resection **AND** Diagnosis: Glioblastoma **THEN** probability of Survival > 1 year : 5.7% (1.6%-12.2%)

⋮

ELSE probability of Survival > 1 year : 47.9% (41.3%-54.6%)

Figure 2.1: An example BRL created from our dataset to predict one year survival. The 95% credible interval of the survival probability is shown in parenthesis.

2.3.1 Frequent Itemset Mining

BRLs are constructed from pre-mined association rules extracted from the dataset. Each rule is composed of two different sets of items, a and b, also known as itemsets. An association 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 first rule in Figure 2.1, the antecedant a is "Diagnosis: meningioma AND Treatment: Excision surgery", and the consequent b is "Survival $>$ 1 Year". Thus if a is true, then the probability of b, or probability of survival greater than one year, is 96%.

Frequent itemset mining [33] is used to find common patterns from the data. Itemsets, consisting of these frequent patterns, are used to construct the association rules. Most rule-mining approaches make the restrictive assumption that all features are binary or categorical. During preprocessing, continuous data (e.g. Age) must be discretised into categorical features using interpretable thresholds (e.g. ages 50-59, 60-69, etc.) or other discretisation methods (see Section 2.6). The BRL model uses the FP-Growth algorithm [34] for frequent itemset mining. For binary and categorical features, the chosen rule mining algorithm does not matter [27]. The rule mining algorithms all perform breadth-first search thus alternative algorithms, such as Eclat or Apriori [35, 36], would generate an identical list of itemsets given the same constraints.

FP-Growth takes a discrete dataset and returns a data subset as itemsets that satisfy constraints on minimum support and confidence. Support, defined in Equation 2.2, refers to how frequently the itemsets appear in the data, where $freq(a,b)$ is the frequency of the itemsets containing items a and b, and N is the number of observations in the dataset. Confidence, defined in Equation 2.3, is the frequency of itemsets that contain a which also contain b, or how often a rule is found to be true [34]. Consequently, performance of these rule mining algorithms is dependent on user-specified thresholds. If the minimum support 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 feasible for effective use. Note that these equations are provided for completeness but they will not be used for rule evaluation and this thesis will focus on standard metrics for machine learning (see Section 4.5).

$$\text{Support} = \frac{freq(a,b)}{N} \tag{2.2}$$

$$\text{Confidence} = \frac{freq(a,b)}{freq(a)} \tag{2.3}$$

2.3.2 Learning Bayesian Rule Lists

The previous section described how frequent patterns, or association rules, are mined from the dataset. The BRL algorithm optimises over these pre-mined rules, rather than the entire feature space, which greatly reduces computation time [27].

The Bayesian approach to building a rule list takes into account user specified priors, which are often used to favor concise rule lists with small rule cardinalities. BRLs create a posterior distribution over rule lists, given the observed data and prior assumptions. Using the generative model outlined in Algorithm 1, the model samples an initial decision list from the posterior distribution that maximises the posterior probability.

The generative BRL model creates a decision list d, for the target labels $\mathbf{y}$, from the features $\mathbf{x}$ and antecedents $\mathbf{A}$. The predicted label $\mathbf{y}$, follows a multinomial distribution over labels (rather than a single label). The multinomial probability is given a prior α which represents the prior pseudo-count for the target classes. The generative model (see Algorithm 1), defines $a_{<j}$ as the antecedents before j in the rule list (if any), c_j as the cardinality of a_j and $c_{<j}$ as the cardinalities before j in the rule list.

Algorithm 1: Generative BRL Model [27]

Result: Sample an initial rule list from the posterior distribution over antecedent lists.

Sample a decision list length $m \sim p(m|\lambda)$;
Sample the default rule parameter $\theta_0 \sim Dirichlet(\alpha)$;
for *decision list rule j = 1, ..., m* **do**
    Sample the cardinality of antecedent a_j in d as $c_j \sim \mathrm{p}(\mathrm{c}_j|c_{<j}, \mathbf{A}, \eta)$;
    Sample a_j of cardinality c_j from $p(a_j|a_{<j}, c_j, \mathbf{A})$;
    Sample rule consequent parameter $\theta_0 \sim Dirichlet(\alpha)$;
end
for *observation i = 1,..., n* **do**
    Find the antecedent a_j in d that is the first that applies to x_i;
    If no antecedents in d apply, set $j = 0$;
    Sample $y_i \sim Multinomial(\theta_j)$;
end

We will now discuss the posterior, prior and likelihood of the BRL model followed by descriptions of posterior sampling and rule list selection.

Posterior: The goal is to optimise the posterior distribution to obtain the best rule list, where the posterior is proportional to the product of the likelihood and the prior. The full posterior model is defined as:

$$p(d|\mathbf{x},\mathbf{y},\mathbf{A},\alpha,\lambda,\eta) \propto p(\mathbf{y}|\mathbf{x},d,\alpha)p(d|\mathbf{A},\lambda,\eta) \tag{2.4}$$

where λ denotes the prior expected length of the decision list and η denotes the prior expected cardinality of a rule. The user must specify the prior hyperparameters: α,λ,η.

Prior: The prior describes the probability of the decision list given the antecedents and specified hyperparameters. The prior is defined as:

$$p(d|\mathbf{A},\lambda,\eta) \tag{2.5}$$

where λ is a hyperparameter for $p(m|\mathbf{A},\lambda)$, and the number of rules m follows a Poisson distribution, truncated at the total number of pre-mined antecedents. The second term η, also follows a truncated Poisson distribution, where values are removed when no rules are available with the specified cardinality. The truncated Poisson distribution is a proper prior and was chosen for its simple parameterisation [27]. The prior multiplicatively combines the distributions from which a decision list d is sampled.

Likelihood: The likelihood of the model describes the probability of the target $\mathbf{y}$, given the features $\mathbf{x}$, decision list d and prior α:

$$p(\mathbf{y}|\mathbf{x},d,\alpha) \tag{2.6}$$

The likelihood increases when the decision list d better describes the data. α, the prior class pseudo-count, is often set to 1 for both classes resulting in a uniform prior.

Markov chain Monte Carlo (MCMC) sampling: The optimal decision list d^* cannot be directly calculated from the posterior distribution. An initial decision list is selected (using Algorithm 1), and iteratively modified using MCMC sampling [37] to generate many samples of decision lists from the posterior distribution. The initial decision list is modified through the swapping, addition, and removal of rules where the modified rules and their new positions are chosen uniformly at random. At each modification, the posterior probability of the decision list is evaluated. For every MCMC iteration, 3 chains are run until convergence (diagnostic $\hat{R} < 1.05$), and each chain was initialised independently from a random sample from the prior. This procedure ensured a variety of lists were produced that are not dependent on one initial decision list.

Rule list selection: Given the posterior distribution, new observations $\widetilde{\mathbf{y}}$ 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 *BRL-point* [27]. Alternatively, the entire posterior can be used to estimate a prediction. Thus the full collection of posterior samples is used to classify $\widetilde{\mathbf{y}}$. This method is called *BRL-post*. By using the entire posterior prediction the classifier is no longer interpretable. One solution is to provide several point estimates from the posterior as example explanations [27]. However, due to its interpretability limitations, we chose not to explore this method.

2.4 Falling Rule Lists

An extension of the BRL algorithm is the FRL model. A FRL is an ordered decision list whereby the estimated probability of success, or $P(Y = 1 \mid X)$, decreases down the list [28]. Similar to BRL, pre-mined itemsets are first extracted from the data, then Bayesian modelling is used to produce a decision list. FRL's use Monte Carlo sampling [37] and simulated annealing [38] to approximate the FRL.

The parameters of the FRL model are as follows: the size of the rule list L, the *if* clauses (or antedecents $\mathbf{A}$), and the risk scores r_l associated with each rule. The risk scores are passed through a logistic function to produce a risk probability between 0 and 1 (i.e. $P(Y = 1 \mid X)$). The rule at the top of the list will have the highest risk score, which will monotonically decrease down the list. These monotonicity constraints are enforced through reparametrization so that $r_l \geq r_{l-1}$for $l = 0, ..., L$. After reparametrization, the rule list prior is:

$$p(d|\mathbf{A}, \lambda, \eta, \gamma, K) \tag{2.7}$$

where γ is used to determine the risk score and K determines the risk score associated with the default rule r_L (see Wang and Rudin [28] for mathematical details). γ follows a truncated Gamma distribution [39], which allows for posterior sampling while enforcing monotonicity constraints. K is also Gamma distributed.

Similar to BRL, to obtain the optimal decision list d^*, Monte Carlo sampling from the posterior distribution is required to generate an initial sample of d. A combination of Gibbs and Metropolis-Hastings sampling [40] is used over the prior parameters in Equation 2.7. Unlike BRL, following the production of an initial rule list, instead of yielding many sample lists, a point estimate of d^* is found using simulated annealing.

Simulated annealing is a global optimization technique that begins with a random initial solution and incrementally improves an objective function [38]. An initial decision list (state s_t) is selected using Monte Carlo sampling and the next state s_{t+1} is

obtained by choosing uniformly at random an operation (swap, replace, add, remove) that alters the current rule list. This new rule list s_{t+1} is accepted if based on an objective function it is better than the current state s_t. If the new state s_{t+1} is not superior, it may still be accepted with a probability specified by a temperature schedule. This temperature parameter slowly decreases thus accepting less bad moves as the algorithm progresses. In the original paper, stimulated annealing was run for 5000 steps then a final rule list, which is an approximate of the global optimum, or maximum posterior probability, was returned [28].

2.5 Baseline and Machine Learning Approaches

Statistical models are commonly used in clinical practice for survival analysis and treatment decisions. The Cox proportional hazards model, introduced in 1972, is a well-recognised statistical technique frequently used in clinical trials to identify differences in survival due to treatment and prognostic factors [41]. Nowadays, with the accumulation of clinical, genetic, and imaging data, these statistical models are often outperformed by ML models. However, statistical methods have proven effective in clinical practice and the Cox model is still wildly used in the medical field today [42].

This project will compare BRL and FRL algorithms to the baseline Cox model and popular ML approaches. To represent the class of uninterpretable ML methods, we chose random forest, logistic regression and support vector machine (SVM). We briefly review the baseline Cox model and various ML approaches next.

2.5.1 Baseline Cox proportional hazards model

The Cox model is a regression method for survival data analysis [41]. The model analyses the effect of several prognostic variables on survival using a hazard function $\lambda(t)$. The hazard function describes the probability a person will experience an event (e.g. death) as a function of time. The model uses a semi-parametric approach whereby the effect of the covariates on $\lambda(t)$ is assumed parametric, but there is no assumption regarding the shape of $\lambda(t)$. Thus the form of $\lambda(t)$ does not need to be specified. The Cox model is defined as:

$$\lambda(t) = \lambda_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k) \tag{2.8}$$

where $\lambda(t)$ is the hazard function for an observation, $\lambda_0(t)$ is the baseline hazard function, $x_1, ...x_k$ are the covariates and $\beta_1, ...\beta_k$ are the coefficients estimated from the

data. The following assumptions are made by the model: each feature independently affects the hazard, and each feature has a multiplicative effect on the hazard function that remains constant over time (see Section 4.2 for a Cox model implementation).

2.5.2 Popular Machine Learning Models

Random Forest: Decision trees, which are the building blocks for random forests, follow a flowchart-like structure that recursively partitions the feature space according to cutoff feature values [43]. Each split results in different data subsets, and each data point belongs to one subset. The final subsets are called terminal or leaf nodes, where the average outcome of the data in that node is used to predict the model output.

A random forest consists of a large number of decision trees that work together to make a prediction [44]. For classification, each decision tree returns a class prediction and the majority response becomes the final prediction. Random forests are interpretable to an extent, where the importance of each feature can be measured using an impurity-based metric [45] (see Section 5.2.2).

Logistic Regression: Logistic regression is a supervised classification technique that models the relationship between input variables and a binary target output [46]. The model fits a sigmoid function to the data and returns a probability of the target output belonging to one of the two classes. Logistic regression is also a partially interpretable model where the importance of each feature is determined using the odds ratio [47] (see Section 5.2.2).

Support Vector Machine: A support vector machine (SVM) is a supervised classification model that separates two classes using a hyperplane [48]. If the classes cannot be linearly separated, an external technique called the kernel trick is implemented to transform the original data into a new feature space [49]. The transformed data can now be linearly separated. The implementation of a kernel results in the complex transformation of data thus rendering an SVM uninterpretable.

2.6 Discretisation

As discussed in Section 2.3.1, rule mining algorithms require features to be binary or categorical, thus continuous features must first be discretised. The chosen discretisation method and splitting criteria can have a significant impact on the final algorithm performance [50]. Ideally, discretisation should partition the feature in a way that

reflects the original distribution. The remainder of the section will introduce discretisation methods used in Section 3.3.5.

2.6.1 Unsupervised Discretisation

Unsupervised discretisation does not require class labels. Thus the discretised data may be used for multiple purposes allowing for more versatility in the application of the final algorithm. Unsupervised discretisation methods include:

Equal Width Binning: This method divides the continuous feature into k bins of equal size. The width of each bin is defined as: $w = (max - min)/k$. For example, the feature *Age* may be split into 10 bins (0-9, 10-19, 20-29, etc.).

Quantiles: Quantiles are a set of 'cut points' that divide a feature into equal-sized, adjacent, subgroups (or divides a probability distribution into intervals with equal probabilities). There is one fewer quantile than number of groups created.

2.6.2 Supervised Discretisation

Supervised discretisation uses information from class labels for optimisation. This method tends to yield higher accuracy, but data is limited to a single application [50]. An extension of the BRL model employed the following method (see Section 3.3.5):

Entropy: Entropy-based discretisation partitions the dataset into intervals that maximises the information from the data, measured using entropy. Entropy is a measure of uncertainty, thus the goal is to reduce the uncertainty in the data (or maximise information gain). Entropy is also used by decision trees to decide which features are used to partition the data. Mathematically, entropy can be written as: H(X) = $-\sum_{i=1}^{n} p(x_i)\log_2 p(x_i)$ where the possible outcomes of the variable X occur with probability $p(x_1)$,...,$p(x_n)$. Information gain (IG) can be written as: $IG(X,Y) = H(X) - H(X|Y)$ where the conditional entropy of X given Y is subtracted from the entropy of X. This returns the information gained about X given additional information about Y. For discretisation, the feature split that results in the highest IG is selected. IG is used to recursively define the best bins, until a stopping threshold is reached. A common criterion is the Minimum Description Length (MDL) principle [51], whereby a feature is recursively split until the IG is lower than the MDL of the split. However, if the MDL method cannot find a way to initially bin the data, it assigns all data to a uniform bin. According to the algorithm, the splitting of the feature did not provide enough information relating to the target class, and the IG is below the MDL threshold.

2.6.3 Heuristic Discretisation

User-specified boundary points are chosen based on an in-depth understanding of the data. To make appropriate discretisation decisions, domain knowledge is often required to understand the feature's behavior and distribution. For example, the feature *Age* has a natural separation (e.g. ages 0-18, 19-30, etc.) but the feature *Maximum Tumour Size* may require consultation with an expert to determine an organic split.

2.7 Existing Work on Brain Tumour Survival Prediction

Now that we have reviewed some of the algorithms employed in the project, let us turn to a discussion of current work on brain tumour survival prediction.

Every day clinicians attempt to tailor medical treatment decisions to patients by taking into account demographics, medical history, and socio-economic factors. More recently, the integration of ML into patient care has re-shaped the way personalised medicine is viewed. ML models have already shown great promise in precision medicine for lung cancer [52], breast cancer [53, 54], and leukemia [55]. Such studies often use genetic or magnetic resonance imaging (MRI) data, with complex black box models to make medical decisions. However, these models are of little use when only clinical data is available. In the interest of brevity, only studies using clinical data will be discussed as they are most applicable to this project. At present, the use of clinical data alone for brain tumour survival prediction often employs statistical techniques and the use of ML is minimal.

Commonly, clinical features are used by survival studies for the development of a nomogram, a popular tool to estimate individualised cancer prognosis [56]. Based on multiple variables, nomograms compute a cumulative point score that is used to predict the probability of an event, such as survival. Barnholtz et al. [57], developed a nomogram to predict the 6- and 12-month survival of patients with a brain metastases. The authors compared three statistical models (Cox proportional hazards regression, recursive partitioning analysis, random survival forests) and found that the Cox model performed the best. This model was then used to build the nomogram. Many of the features included in the nomogram were similar to the ones assessed in this project, such as: age, Karnofsky Performance Scale (or KP score), surgery type and patient self-reported symptoms. Interestingly, of all the variables used, KP score, which is commonly used in oncology to assess the functional state of a patient (see Section

3.3.3), had the smallest effect on the incremental improvement of accuracy [57].

Gittleman et al. [58], also created a nomogram to assess the 6-, 12-, and 24-month survival probability of patients with a glioblastoma, the most common type of brain tumour. The authors compared the same three statistical models and again found that the Cox model outperformed the rest. An older age at diagnosis, male sex, lower KP score and not undergoing a total tumour resection were found to decrease the probability of a longer survival. The observation that men are more at risk than women is in-line with current literature and our own dataset. The literature states that the incidence rate of a glioblastoma in males is 1.6 times greater than in females [59]. In our dataset, of the 519 male patients, almost 45% have a glioblastoma (227), compared to 30% of the 518 female patients. Nomograms have also been developed to predict survival probability for lower-grade gliomas (LGG) [60], and subsequent brain metastasis following metastatic breast cancer [61].

Although nomograms are useful tools for optimising treatment approaches, these models are built using simple multivariable regression techniques, such as the Cox model. However, with the increased availability of clinical data, ML can be used to build predictive models with superior performance and generalisability [62, 63].

2.7.1 Previous Masters Dissertation

This project built upon a previous Master's dissertation that investigated influential factors for brain tumour survival [64]. The previous dissertation used a subset of the same dataset as this project, focussing only on patients with a glioblastoma (see Section 3.1 for dataset description). Survival was treated as continuous, and the limitations of the Cox model were explored. The author compared different tree-based models to the Cox model, and using the average mean square error found that random forest performed the best. To understand what influences survival, the importance of each feature in the random forest model was analysed. The author found that the amount of temozolomide, a type of chemotherapy drug, and the total dose of radiotherapy were most influential (these variables were not readily available in our dataset). This was followed by the extent of tumour resection and age. The author also explored different methods of data imputation and whether current treatment protocols were optimal for patient survival. This project will utilise these imputation techniques and compare additional ML models (logistic regression, SVM) as well as interpretable ML models (BRL, FRL) to the Cox and random forest model (see Chapter 5 for model results).

Chapter 3

Data Preprocessing

Data analysis for this project was carried out in Jupyter Notebooks [65] and code was written in Python 2.7.18 and Python 3.6.10. Standard data analysis libraries were used (Pandas [66], Numpy [67], Matplotlib [68], Scikit-learn [69]). Code for both the BRL and FRL methods were adapted from the original authors' code [27, 28][1].

3.1 REDCap Dataset

This project used the anonymised REDCap Dataset collected by Dr. Paul Brennan. This contains 1391 patient records and 225 feature types for each patient. An earlier version of this dataset was used by the previous Masters dissertation described in Section 2.7.1 [64]. The dataset has since been updated and now includes additional patient cases, more detailed records and updated survival times. Features in the dataset include patient demographics (e.g. sex, age), medical history (e.g. history of cancer, comorbidity), symptom features (e.g. symptom types, duration), radiological tumour analysis (e.g. type, size, location) and treatment details (e.g. type of surgery, extent of resection). Survival time for each patient is recorded up to April 2020. A preliminary analysis of the dataset was carried out to assess feature distributions, relationships between features, and the extent of missing data.

3.2 Exploratory Data Analysis and Missing Data

An initial review of the dataset revealed that a large selection of features relating to symptoms and signs were duplicated (see Section 3.3.1 for handling of duplication).

[1] https://users.cs.duke.edu/∼cynthia/papers.html

Additionally, a significant number of features related to blood test results (pre- and post-chemotherapy) and other in-depth treatment details. For the purposes of this project, these features were ignored and a subset of 21 features were selected for further investigation (see the Glossary in the Appendix for features descriptions).

Our preliminary investigation found that the dataset contained 111 records of patients who either did not have a brain tumour or did not receive a diagnosis. Alternative diagnoses included stroke, hematoma and atrophy (cell degeneration). These 111 patients were removed from the dataset. Additionally, a large section of patients (243) were missing a significant number of diagnostic features, including presenting symptoms and signs. Dr. Brennan advised that the dataset was compiled by different authors at different time points thus some authors did not record certain features. These patients were removed from analysis to improve the richness of the dataset. The following statistics are based on the remaining 1037 patients.

The dataset contains significant heterogeneity with more than 20 different brain tumour types. The most common types are glioblastoma (381), metastasis (189), meningioma (170) and glioma (109). The minimum age of a patient is 16 years while the oldest is 97 years. The majority of patients are aged 50-70 with a median age of 61 years. There are an equal number of male and female patients (519 and 518, respectively). 18% (189) of patients had a previous history of cancer and 47% (491) of patients presented with a co-morbidity, the most common being cardiovascular (18% of all patients). The most common location for a tumour was in the frontal lobe (35%) followed by the temporal lobe (23%). More than half of the patients (64%) had some type of surgery, 25% of patients underwent chemotherapy, and 20% received radiotherapy. This is inline with current treatment protocols where surgery is often the first line of treatment, followed by radiotherapy and concurrent chemotherapy for more advanced tumours [70]. Finally, 22% of patients received no treatment.

A large portion of the dataset pertained to symptom and sign features. The main difference between symptoms and signs is who observes the effect. A symptom is observed by the patient themselves (subjective) while a sign is observed by the physician (objective). Symptoms are often what drive a patient to consult a physician. The three most common symptoms patients presented with were headache (295), unilateral weakness/change (89) and seizures (85). The three most common signs patients presented with were unilateral weakness (125), cognitive/non-specific confusion (99) and problems walking/ataxia (84). Although all patients in the reduced dataset presented with at least one symptom, 41% of patients did not present with any signs.

Patient survival was measured in days and 35% of patients were still alive (361) (last updated April 2020). 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 [71]. The mean survival time is 383 days, while the median is 245 days. Overall survival time follows a decreasing exponential distribution (see Figure 3.1). Currently, the largest survival time is 3964 days, or about 10 years, however only three patients have a survival time greater than 2000 days (for sake of visualisation they were not included in Figure 3.1). The high variance in survival may relate to a variety of factors including age, tumour type, and treatment outcomes.

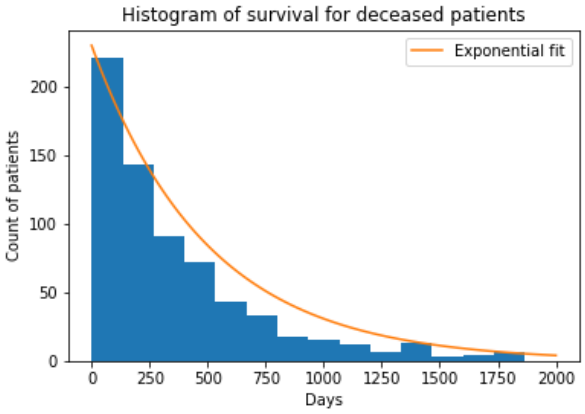

Figure 3.1: Histogram of survival in days. The graph follows a rough exponential shape.

3.2.1 Missing Data

The majority of patients in the reduced dataset have incomplete records. The bulk of features are 70-80% complete, with a mean completeness of 81% and median of 88% (see Figure 3.2). An important consideration is the handling of these missing values. Treating all missing values the same would be a strong oversimplification as missing data can come from a variety of sources. 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. Imputing this missing data would introduce bias into the dataset, and a patient presenting with only one symptom may in itself be informative. Given the amount of missing symptom and sign data, only symptom 1 (i.e. the first symptom a patient presents with), symptom 2 and sign 1 were used as features.

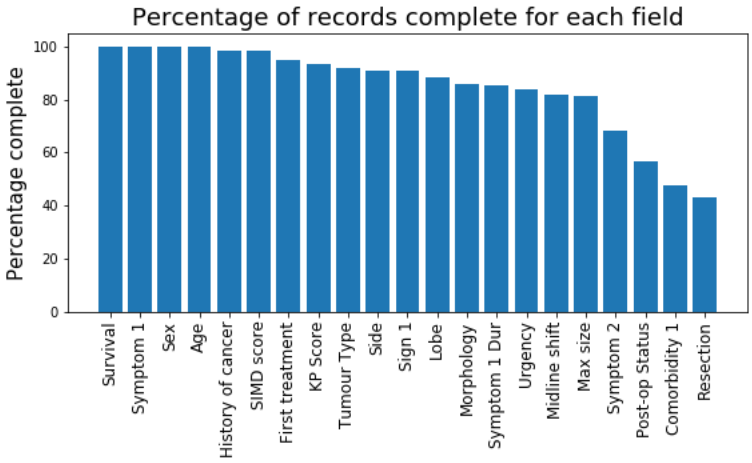

Figure 3.2: Percentage of records complete for each feature. The following features are 100% complete: Survival, Symptom 1, Sex, Age. See Glossary for feature definitions.

3.3 Preprocessing

Due to the amount of incomplete data, the initial dataset was not suitable for survival prediction. A number of data cleaning steps, including imputation and discretisation were first required.

3.3.1 Symptom and Sign Features

The original dataset contained almost 90 features relating to symptoms and signs, which was partially due to feature duplication. First, the symptoms and signs a patient presented with at their general practitioner (GP) were recorded, as well as the symptoms and signs a patient had upon referral to a clinician. The majority of these presenting features were repeated and thus were concatenated into a single feature. The number of presenting symptoms (i.e. symptom 1, symptom 2, etc.) were kept separate.

The symptom and sign data also had a high cardinality. The dataset contained 37 different symptom types and 26 different sign types. Many of these feature types pertained to a small number of patients. This large diversity in patient features may result in an uneven distribution of feature types between the training and test data. After consultation with Dr. Brennan, a decision to group symptom and sign data into larger overarching domains was made. In particular, symptom types were grouped into six domains based on the paper by Ozama et al. (an outline of the symptom groupings are summarised in Table 3.1) [72]. Sign data was grouped into six larger domains based off of groupings directly provided by Dr. Brennan (see Table 6.1 in the Appendix).

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 3.1: Symptom domain classifications based on Ozawa et al. [72], with examples of symptom types in REDCap dataset.

The categorisation of both symptom and sign data created a more homogeneous set of feature types that would prove more informative.

3.3.2 Radiological Diagnosis

Brain tumours have over 130 types that are often named based on the type of cell they develop from or by the location in the brain they originate [73]. As expected, in our dataset the tumour types had a high cardinality (more than 80 types). However, many of these types referred to the same type of tumour (e.g. meningioma suprasellar and meningioma at CP angle). The dataset also contained multiple records of patients who did not have a brain tumour, hence some types referred to alternative diagnoses. The brain tumour types were reorganised and reduced to a cardinality of 10, and types that appeared in less than 10 patients were grouped into a "Rare Tumour" category. See Figure 6.1 in the Appendix for the distribution of brain tumour types in our dataset.

3.3.3 Karnofsky Performance Status

The KP scale is a standard way of assessing a cancer patients ability to perform every-day tasks [74]. The score ranges from 0 to 100 and is commonly scored in deciles. A score of 0 equates to death, while a score of 100 represents an asymptomatic individual with normal function (see Table 6.2 in the Appendix for the original definition of KP scores). The scale is a 'gold standard' in clinical oncology and is commonly used as to determine a patient's ability to tolerate treatments, such as chemotherapy. However, the values of the KP scale are ordinal. This means that a value assigned to a patient is based on a ranking but the numerical value associated with this rank is meaningless.

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. A patient's KP score is often assessed by clinicians, and when compiling a dataset this can result in interobserver subjectivity [75, 76]. Both the ordinal nature and subjectivity of the KP scale is a major limitation.

The previous Master's dissertation treated the KP score as continuous [64]. After a review of the literature and consultation with Dr. Brennan, we decided to treat the KP score as categorical. To reduce the bias associated with the KP score, the values were grouped into three overarching states which describe different levels of performance. The groupings are as followed: A (Score 80-100): normal work and self-care, B (Score 50-70): unable to work but can care for most personal needs, C (Score 0-40): unable to care for self. However, due to the large number of patients in group A (70%), we decided to further break down the KP scores into the following groups: < 50, 50-70, 80, 90, 100. This would allow for a more fine-grained analysis of the data.

3.3.4 Data Imputation

As described in Section 3.2.1, a significant number of patients have incomplete records. Some of these features are correctly missing while others are not. Commonly, missing data is managed by either the deletion or imputation of values. The deletion of data may lead to bias or loss of statistical power, but the imputation of missing values retains all the data which is preferred due to our small dataset size. However, imputation may introduce a different kind of bias, and the chosen method may influence the final results. Imputation techniques developed in the previous Masters dissertation were employed in this project and compared on a model-by-model basis [64].

For categorical variables, the simplest method for imputation is to use the mode, and for continuous variables, the analogous approach is the mean. For cases where outliers may affect the mean, the median may be used. However, the continuous variables in our dataset are constrained thus outliers are not a concern (e.g. the feature *Symptom 1 Duration* must fall between 0 and 52 weeks).

A more complex imputation approach is k-nearest neighbours (k-NN) [77]. This approach finds the k nearest neighbours (based on a distance metric) for a missing observation from all complete observations in the data. The missing observation is replaced with the most frequent value (for categorical variables) or mean value (for continuous variables) from the neighbouring observations. The use of normalised and

unnormalised features were both explored [64]. The purpose of normalisation is to change the numerical feature values in a dataset to a common scale [78]. The distance metric for k-NN uses feature values, thus if one distance is larger than the other, that feature will dominate the k-NN outcome.

A final method of imputation is to fill in the missing values using information from other variables by means of a regression model [79]. A regression model is estimated using the observed data and the fitted regression weights are used to predict the missing values. This approach assumes that the variables are not independent, hence other variables can be used for imputation. For continuous variables, a linear regression model was used, and for binary variables, a logistic regression model. Logistic regression was extended to multi-class variables by using a 'one-vs-rest' scheme [80].

The goal of each model was to individually predict the missing values. The mode- and mean-fill was used as baseline methods for categorical and continuous imputation. For a given feature, the k-NN and regression models were trained using data points with features that were at least 40% filled (this value was determined by the previous dissertation) [64]. For each feature 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, and the results are shown in Table 6.3 and 6.4 in the Appendix. The optimal imputation method for each feature was then implemented on the full dataset, and the missing variables were replaced with the model's predicted output.

Our results are similar to that of the previous Masters dissertation [64]. For categorical variables, logistic regression often performed the best. Compared to k-NN, logistic regression learns the weighting (or importance) of variables in the data, hence it is more likely to make accurate predictions. k-NN does not learn patterns from the data but assumes that similar variables exist in close proximity. As expected, k-NN with normalised features often outperforms regular k-NN. As k-NN uses distance to learn, this highlights the need for feature values to be on the same scale for optimal performance. For continuous features the results were similar, whereby all methods performed better than baseline and normalised k-NN outperformed regular k-NN.

3.3.5 Discretisation

As discussed in Section 2.3.1, association rule-mining requires categorical features. Discretisation can be used to improve the clarity of classification models by extracting useful intervals that occur at discontinuous regions of a feature's distribution.

We compared three discretisation methods: binning, quantiles and heuristic discretisation. These methods were assessed using the classification accuracy, F1 score and area under the receiver operating characteristic (AUROC) curve, of the discretised features compared to the continuous features, using the BRL model (see Section 4.5.1 for description of metrics). The BRL default hyperparameters were used, and patient survival greater than one year was evaluated on the entire dataset using 10-fold-cross validation. Other methods for discretisation, including k-NN and decision trees, were also attempted, but due to time constraints they were not investigated further. Additionally, an extension of the BRL algorithm, which employed a discretiser to handle continuous data, was explored[2]. The algorithm used the MDL principle criterion (introdcued in Section 2.6). However, the discretiser was unable to adequately partition the feature space, and the variables were assigned to a uniform bin. Thus a decision was made to discretise all continuous data before model training.

The features that underwent discretisation were 'Age', 'Symptom 1 Duration' and 'Maximum Tumour Size'. Each feature was divided into bins (and quantiles) 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) (see Figure 6.2 in Appendix for each feature's distributions). Note that the feature *Symptom 1 Duration* followed a strong bimodal distribution resulting in errors for quantiles above 4. A natural (heuristic) discretisation of the data was also investigated. For example, the feature *Age* has an inherent separation (e.g. children, youth, adults, seniors). Dr. Brennan was also consulted for advice on the heuristic discretisation of these features.

The prediction accuracy and AUROC as a function of the number of bins and quantiles is illustrated in Figure 6.3 in the Appendix. The discretisation results are shown in Table 6.5 in the Appendix. Based on these findings, Age and Maximum Tumour Size were discretised manually, and Symptom 1 Duration was discretised using 6 bins.

[2]https://github.com/csinva/interpretability-implementations-demos/tree/master/imodels/bayesianrulelist

Chapter 4

Methods

As stated in the introduction, our main objective was to compare the performance of interpretable rule list models to popular ML algorithms for the prediction of brain tumour survival. Random forests, logistic regression and SVM classifiers were chosen as alternative ML models for both their popularity and relative lack of interpretability (see Section 2.5 for model descriptions). Random forests and logistic regression provide means by which the model's feature importance can be ranked and thus are interpretable to an extent. Due to the non-linearity of SVMs, feature importance cannot be directly determined, hence they are classified as black box models. Artificial neural networks are often viewed as the quintessential black box ML model due to their multilayer nonlinear structure [81]. However, neural networks require large datasets for training, thus due to the small size of our dataset neural networks were not explored as a type of black box ML model.

Recall that one of the original goals of this project was to develop a rule list classifier for brain tumour diagnosis. We expected to acquire an additional dataset, called Clinical Practice Research Datalink (CPRD), that would augment the REDCap dataset. The CPRD dataset contained over 30,000 primary care medical records from both healthy controls and brain tumour cases. However, this dataset could not be secured in time (see Section 6.1 for discussion on future work). Nevertheless, we did briefly look at the implementation of rule lists for the classification of a glioblastoma versus other cancer types using the (relatively limited) information that was available in the REDCap dataset. Glioblastomas were chosen as it was the most common tumour type in the dataset and is also the most fatal [4]. See Section 5.3 for a discussion.

4.1 Data Preparation

The initial cleaning and preprocessing of the dataset is described in Chapter 3. The final dataset contained 1037 patients with 21 clinical features. Missing variables were handled in one of two ways. For categorical variables that were truly missing (missing on purpose), the values were assigned to a single null group. For example, if a patient was missing a value for the feature Sign 1, the null value was assigned to the group 'no signs'. For categorical and continuous variables that were not truly missing (missing by accident), imputation was performed (as described in Section 3.3.4). To support the rule list models, all features types were categorical. One-hot-encoding was used for models that were unable to handle categorical features. One-hot-encoding, also called dummy coding, creates a separate binary feature that takes the value 0 or 1 to indicate the absence or presence of a categorical value.

The final dataset was separated into a training set (80% of the data) and test set (20% of the data). The hyperparameters for each model were fine-tuned using grid search with 5-fold-cross validation across the training set. Based on the optimal hyperparameters, the model was re-trained on the entire training set and evaluated using the held-out test set. 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: 440 patients (42%) survived less than a year and 597 patients (58%) survived greater than a year. Patients who were alive at the time of data collection were included in the survival greater than a year group. Glioblastoma labels were created from the diagnosis data resulting in the following data split: 381 (37%) patients had a glioblastoma and 656 (63%) patients did not have a glioblastoma. Model performance was assessed using the metrics described in Section 4.5.1.

4.2 Cox Model

The Cox model was implemented using the lifelines python package[1] and served as an alternative baseline for the ML models. The Cox model requires numerical features and 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. A penalizer was added to the model which reduces the size of the coefficients during

[1] https://lifelines.readthedocs.io/

regression thus controlling for high correlation and improving the stability of the estimates [82]. A search for the optimal penalty parameter between values of 0.01 and 0.3 was performed using 5-fold-cross validation and a value of 0.2 was selected.

The purpose of the Cox model is to predict survival time thus the model was trained using the continuous survival data. The predicted survival time was then converted to binary one year survival labels and these labels were used to calculate the model's performance. The feature importance was calculated directly from the Cox model using the variable's coefficient and standard error value. The p-values can be calculated for the null hypotheses that the coefficient value is 0 (i.e. the feature has no importance).

4.3 Rule Lists

The rule list models are dependent on hyperparameters for both the FP-Growth algorithm and the decision list priors. For both models, α, the prior pseudo-count for the classes (see Section 2.3.2), was set to [1,1] for simplicity, resulting in a uniform prior. An exhaustive grid search over the following hyperparameters was performed: the minimum support threshold for the rules, minimum and maximum rule cardinality, and length of the rule list.

Typically minimum support threshold is set to 10% [33], thus the range of values searched was 5%, 10% and 15%. Rule cardinality was selected on the basis of Dr. Brennan's council that interpretations are more difficult with high cardinality rules because it is "harder to reconcile the combinations of features as being clinically logical". Thus a minimum rule cardinality of 1 and 2, and maximum rule cardinality of 2 and 3 were explored. Finally, the rule list expected length values was selected based on the original papers' parameters [27, 28]. The values searched were: 5, 8, 10, 12, 15.

Bayesian Rule Lists: Based on the hyperparameter tuning with grid search, the optimal BRL hyperparameters included rules with a minimum support of 5%, a minimum rule cardinality of 2 and a maximum cardinality of 3, and a rule list with a prior expected length of 8. The number of antecedents used ranged from 2817 to 2923 across the training folds. The number of MCMC chains was set to a default value of 3, and the algorithm was run for a maximum of 10,000 iterations. Using these hyperparameters, the BRL model was fitted on the full training dataset and the BRL point estimate was used to evaluate model performance on the test set.

Falling Rule Lists: For the FRL model, rules with a minimum support of 5% were mined, and a minimum and maximum cardinality of 2 conditions per rule was

chosen. The prior expected rule list length was set to 10. Additional hyperparameters for the simulated annealing algorithm were selected based on the FRL paper's original parameters [28]. A default temperature of 1 was chosen for simulated annealing and the algorithm was run for 5000 steps. With these hyperparameters, the FRL model was retrained on the entire training set, and the optimal rule list returned by simulated annealing was used on the test set to evaluate model performance.

4.4 Machine Learning Models

Random forest, logistic regression and SVM models were all implemented using the scikit-learn package [69]. The three models cannot directly handle categorical variables thus one-hot encoding was implemented for all features. As with the rule list models, a grid search for the model's hyperparameters was carried out using 5-fold-cross-validation across the training dataset.

Random Forests: Random forest performance is conditioned on multiple hyperparameters including the number of trees, the maximum depth of a single tree, the minimum number of observations required at a leaf node and the number of features to consider when looking for the best split. Due to the range of hyperparameters, a set of candidate hyperparameters were chosen based on the previous Masters dissertation [64]. The optimal hyperparameters were found to be the training of at least 400 trees with a maximum depth of 10 and a minimum number of 3 observations at each leaf node. A maximum of 8 features were considered at each split.

Logistic Regression: Logistic regression required the fine-tuning of two hyperparameters, the type of solver for optimization and the regularization parameter C. Regularisation is used to discourage the learning of complex models to prevent overfitting. The strength of regularization is determined by C, where the default value is one and a smaller C value represents stronger regularization. The final logistic regression model used the *lbgfs* solver which supports L2 regularisation [83]. L2 regularisation determines the loss function penalty by using the squared value of the model's weights (coefficients). The regularization strength parameter C was set to 0.1 (see Figure 6.4A in the Appendix for the effect of C on model performance).

Support Vector Machine: For SVM, the type of kernel, regularization parameter C and gamma were tuned. Gamma is a kernel parameter which determines the curvature of the decision boundary. A smaller gamma value signifies less curvature thus the complexity or "shape" of the data is captured less. The final SVM model implemented

the radial basis function (rbf) kernel [84] with a regularization strength of 2 and gamma value of $\frac{1}{k}$ where k is the number of features. Figure 6.4B in the Appendix shows the effect of C on training and validation set performance for the SVM model.

4.5 Evaluation Criteria

Multiple methods of evaluation were used to compare the performance of the classification models. First, standard evaluation metrics were used to assess the models predictive ability and second, the model was evaluated for its level of interpretability.

4.5.1 Standard Evaluation Metrics

The predictive ability of each model was assessed using the following metrics:

Accuracy: Accuracy is a simple metric that measures how often a data point is correctly classified. Accuracy works well when there are an equal number of samples in each class. For highly imbalanced classes, a model may achieve a high accuracy without actually leaning anything from the data [85]. The classes in our dataset are relatively balanced (42% and 58%) thus this is not a major concern.

F1 Score: The F1 score is also a measure of a model's accuracy with values between 0 and 1 [86]. The F1 score is the harmonic mean of the model's precision and recall where a score of 1 represents perfect precision and recall. Precision is the number of correct positive results divided by the total positive results returned by the classifier. Recall is the number of correct positive results divided by the number of samples that should have been identified as positive. A model with high precision but lower recall would be extremely accurate but often misses instances that are difficult to classify. The F1 score is the preferred metric when classes are imbalanced because it gives a better measure of the incorrectly classified cases (unlike accuracy).

Receiver operating characteristic (ROC) curve: The ROC curve plots the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings for a binary classifier [87]. The area under the ROC curve (AUROC) measures the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. AUROC values range from 0 to 1, where a value of 0.5 represents a classifier that performs no better than chance and a value of 1.0 represents a perfect classifier.

4.5.2 Interpretability Metrics

Measuring the interpretability of a model is often difficult due to the subjective nature of the task. Unlike classification performance, there is no standard metric for interpretability that can be used across all models. This makes comparison of model interpretability more challenging. In this project, a model's interpretability will be assessed in one of following ways.

Feature Importance: Unlike rule lists, most ML models do not output a clear set of decision guidelines. Instead, their interpretability can sometimes be assessed by examining the relationships between features that contribute to classification. This is a type of post-hoc interpretability analysis whereby feature importance is analysed after model construction. This analysis will provide an overview of what features the model favors when making predictions, but unlike rule lists, it does not provide fine-grained details on how individual instances are classified.

Local Surrogate Models: Black box models can be assessed for interpretability by using local surrogate models (introduced in Section 2.1.1). This project briefly explores the use of LIME [18] to examine individual predictions from the ML algorithms. The features used to make an individual prediction can be compared to the features used by the rules lists and features found important for global approximations.

Qualitative Evaluation: Empirical evidence of interpretability can be obtained by letting domain experts judge the understandability of a model's output. In this project, Dr. Brennan and his colleague Michael Poon were consulted for their clinical expertise in the field of neurology and brain cancer. Mr. Poon is a neurosurgical trainee in Edinburgh and is undertaking a Cancer Research clinical PhD fellowship with the Brain Tumour Centre of Excellence initiative funded by Cancer Research UK.

The experts were given multiple BRL and FRL point-estimates for analysis. Using the final rule list models, several point-estimates were generated by running 10-fold cross validation on the reduced dataset. The decision lists with the highest AUROC values were selected. A total of three BRLs and one FRL for survival prediction, and one BRL and one FRL for glioblastoma prediction were provided (see Appendix B for all the rule lists). Dr. Brennan and Mr. Poon were asked to consider if the rules produced were sensible, if any rules were surprising or unrealistic, and the potential employability of such a model in a clinical setting.

Chapter 5

Results and Discussion

5.1 Model Performance

This project compared six different approaches for the prediction of brain tumour survival one year after diagnosis. The two rule lists were compared to the baseline Cox model and three ML models that varied in their interpretability level. Each model was evaluated using the metrics outlined in Section 4.5.1. The model's performance is summarised in Table 5.1 and the ROC curve for each model is visualised in Figure 6.5 in the Appendix. Note the ROC curve requires the probability estimates for each class, which could not be directly computed for the Cox model.

	Cox	BRL	FRL	RF	LR	SVM
Accuracy	0.8077	0.8212	0.7927	0.8413	**0.8462**	0.8365
F1 Score	0.8075	0.8143	0.7918	**0.8460**	0.8450	0.8350
AUROC	0.8044	0.8105	0.7814	0.8365	**0.8440**	0.8338

Table 5.1: Performance metrics were assessed on the test set. For each metric, the best model is highlighted in bold. (RF = random forest, LR = logistic regression)

Surprisingly, the Cox model performed on-par with most models and outperformed the FRL model. The FRL algorithm performed the worse of all models. This is not surprising as the model's strong monotonicity constraints may sacrifice performance [28]. However the loss in model performance is minimal, and the level of interpretability associated with this model may be favored amongst medical experts. The BRL algorithm outperformed the baseline Cox model and FRL model, and was comparable to the three ML model's performance. Our results are similar to that of the original

BRL paper, which found the BRL point estimate to perform on par with random forest, logistic regression and SVM [27]. The authors found that BRL-post (introduced in Section 2.3.2) matched random forest for the best performing model. Future work may be undertaken to establish whether BRL-post is a strong competitor for ML models.

Figure 5.1 and 5.2 shows a point-estimate obtained from training the BRL and FRL models. For both rule lists, once a patient has satisfied a rule they will not be taken into account by the rules further down the list. The final rule will only consider a subset of patients that were not classified by the previous rules.

IF Urgency of referral: Suspicion of Cancer (within 2 weeks) **AND** KP Score : 50-70 **THEN** probability of Survival > 1 year: 7.1% (0.9%-19.0%)

ELSE IF Diagnosis: Meningioma **AND** PMH of cancer: No **THEN** probability of Survival > 1 year: 96.0% (91.9%-98.7%)

ELSE IF Symptom 2: Other/Non-specific **AND** Morphology: Heterogeneous **THEN** probability of Survival > 1 year: 37.9% (21.5%-55.9%)

ELSE IF Post-op Status: No Surgery **AND** Morphology: Heterogeneous **THEN** probability of Survival > 1 year: 15.0% (9.5%-21.6%)

ELSE IF KP Score: 100 **AND** Midline shift: 0 **THEN** probability of Survival > 1 year: 93.4% (88.4%-97.1%)

ELSE IF Diagnosis: Metastasis **AND** Age: 65+ **THEN** probability of Survival > 1 year: 25.0% (11.1%-42.3%)

ELSE IF Diagnosis: Glioblastoma **AND** Resection: No Resection **THEN** probability of Survival > 1 year: 21.5% (12.5%-32.2%)

ELSE IF Age: 19-50 **AND** PMH of cancer: No **THEN** probability of Survival > 1 year: 89.3% (81.9%-94.9%)

ELSE IF Resection: < 50% **AND** Urgency of referral: Emergency **THEN** probability of Survival > 1 year: 18.2% (5.4%-36.3%)

ELSE probability of Survival > 1 year: 56.7% (50.1%-63.3%)

Figure 5.1: BRL-point estimate. The 95% credible interval is given in parantheses.

IF Morphology: Homogeneous **AND** KP Score: 100 **THEN** probability of Survival > 1 year is 97.37%, *Support*: 152

ELSE IF Diagnosis: Meningioma **AND** PMH of cancer: No **THEN** probability of Survival > 1 year is 90.91%, *Support*: 55

ELSE IF Age: 19-50 **AND** Urgency of referral: Emergency **THEN** probability of Survival > 1 year is 79.05%, *Support*: 105

ELSE IF Post-op Status: 0.0 **AND** Comorbidity 1: NONE **THEN** probability of Survival > 1 year is 73.61%, *Support*: 72

ELSE probability of Survival > 1 year is 32.29%, *Support*: 446

Figure 5.2: FRL-point estimate. The support indicates the number of patients classified by that rule.

An initial glance at the rule lists shows that BRL is much longer than FRL. The prior expected list length of each model was 8 and 10, respectively. The BRL algorithm produced a list with 10 rules while the FRL algorithm produced a list with only 5 rules. Although this prior is taken into consideration, ultimately the rule list that best supports the data is returned. The shorter length of the FRL is likely due to the monotonicity constraints. The goal of our FRL is to predict the probability of survival greater than

a year, thus all rules that favor survival less than a year are summarised into one final rule. The BRL model does not follow any monotonicity constraints, thus rules that favor both survival less than and greater than a year are included. The more fine-grained approach of the BRL is likely why this model outperforms the FRL algorithm.

5.2 Model Interpretability

The interpretability of each model was assessed using the metrics outlined in Section 4.5.2. 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 reduced dataset to assess feature significance. The final models interpretability was then evaluated using feature importance (Cox model, random forest, logistic regression), LIME (SVM, random forest, logistic regression) and qualitative assessment (BRL, FRL).

5.2.1 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 mlxtend[1] was implemented. This method adds or removes one feature at a time based on a classifier performance until a subset of k desired features is reached. The method allows for a range of k-features to be specified and the feature combination that scores the best during cross validation is returned. Due to comatability issues, the rule lists could not be run in conjunction with the feature selector thus random forest was chosen as the classification model. The model was run using its default parameters and features were assessed using AURUC and 5-fold cross validation. Due to one-hot encoding, a total of 110 feature types were evaluated. The selector was given a range of 3 to 50 features, and 16 features were returned (see Table 6.7 in Appendix for the list of features). The selected features included KP score, symptom 1, age and first treatment. Notably, these features were all used by the nomograms introduced in Section 2.7. Additionally, all of the 16 features selected were used by at least one of the trained models.

[1] http://rasbt.github.io/mlxtend/user_guide/feature_selection/SequentialFeatureSelector/

5.2.2 Feature Importance

The Cox, random forest and logistic regression models all provide means for an analysis of feature importance. An examination of the feature weightings can give rise to simple interpretations of how the model made its classifications. Although not fully transparent, this allows moderate insight into how the model works and may assist clinicians in understanding causal factors for patient survival.

The Cox model feature importance was estimated using the reduced dataset and is illustrated in Figure 5.3. 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. The results are as expected with a worse post-operative status (see Table 6.6 in Appendix for post-operative scale), no treatment, older age and a KP Score below 70 signifying poorer survival. Comparatively, chemotherapy, a younger age, and lower post-operative status signifies better survival. A number of tumour types also appeared important for survival time. As expected, due to the severity of glioblastomas and brain metastases, both tumour types indicated poorer survival. On the other hand, low grade gliomas (LGG) and pituitary tumours signify increased survival time. Both tumours are highly treatable, with a 76% [88] and 82% [89] five-year survival rate, respectively (compared to 5% for glioblastoma [4]). Additionally, tumour morphology type was found to be an important indicator of survival. A heterogenous tumour likely signifies reduced survival time as the tumour contains diverse cell types with distinct molecular structure that may have different levels of sensitivity to treatment [90]. Surprisingly, symptom and sign data was not found to be important. These findings are in contradiction with the previous Masters dissertation which found a number of sign and symptom features to be important predictors for the Cox model. However, the previous dissertation did not re-group the symptom and sign data into larger domains, thus the large cardinality may have had an effect on the Cox model performance.

Random forest feature importance is often measured using mean decrease in impurity (MDI) [45]. This metric determines which variables are split during training, and the decrease in impurity of each feature is averaged over all trees in the forest. The higher the value, the more important the feature (i.e. the better the feature is at splitting the data). However, this metric is biased towards features with high cardinality [91]. As an alternative, the permutation importance of the random forest model was computed as shown in Figure 5.4. The permutation importance measures the decrease in a model's performance when a feature value is randomly shuffled [44]. However,

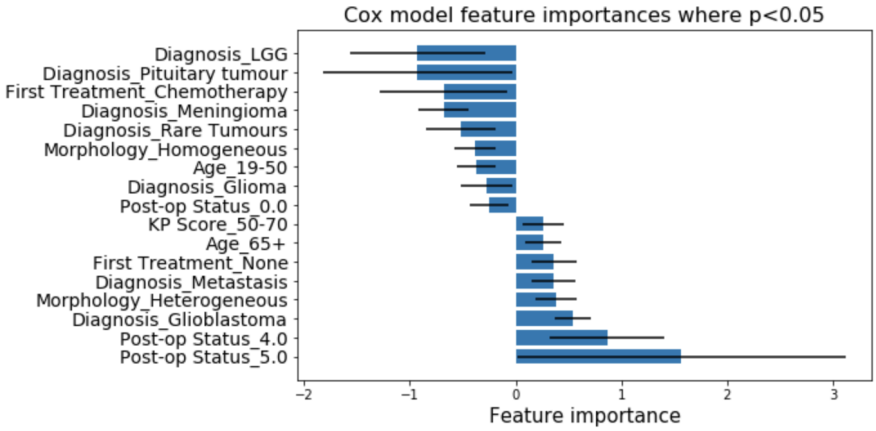

Figure 5.3: Cox model feature importance fitted to the final dataset. All features have a p-value less than 0.05 and the 95% confidence interval are shown with black error bars.

the variable importance does not indicate whether the feature is positively or negatively correlated with survival. The three most influential features were younger age, homogeneous morphology and a glioblastoma tumour (similar to the Cox model).

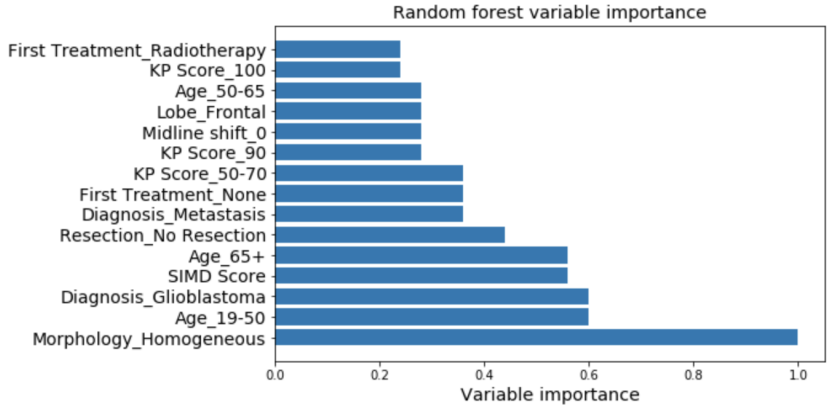

Figure 5.4: The 15 most influential features from the random forest model fitted to the final dataset. Note that the importance does not specify positive or negative correlation.

For the logistic regression model, feature importance was assessed using the odds ratio [47]. The coefficient (or weight) of each feature is equivalent to the natural logarithm of the odds ratio. A feature with an odds ratio closer to one has the least impact, while a higher or lower ratio is more influential. An odds ratio greater than one describes a positive relationship (i.e. increases the odds of survival greater than a year) and an odds ratio smaller than one describes a negative relationship. Figure 5.5 shows the top 8 features with the highest and lowest odds ratio. The features found most influential were similar to the Cox and random forest models. Again, glioblastomas

and brain metastases, heterogenous morphology and older age were strong negative predictors of survival, and meningioma, younger age and lower post-operative score were positive predictors for survival.

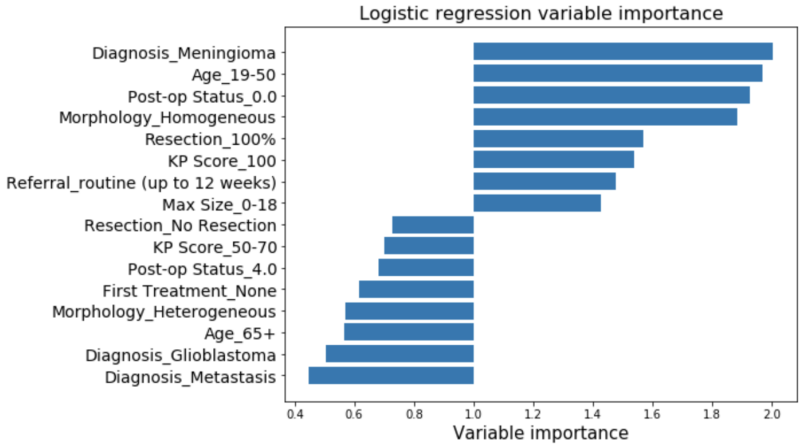

Figure 5.5: The 8 features from the logistic regression model (fitted to the final dataset) with the highest and lowest odds ratio.

5.2.3 LIME

Feature importance could not be directly computed from the SVM model, thus LIME was used as a type of post-hoc interpretability method (described in Section 2.1.1). Compared to feature importance, LIME assesses interpretability at the local level for individual predictions, rather than at the global modular level. LIME was also applied to random forest and logistic regression, and the same prediction instance was compared across the three models. Figure 5.6 shows an explanation of an observation in the test set generated using LIME on the SVM model (see Figure 6.6 and 6.7 in the Appendix for random forest and logistic regression LIME output).

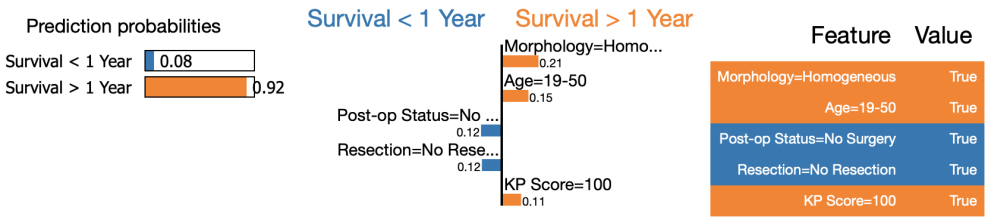

Figure 5.6: SVM feature importance determined by LIME. Negative (blue) features indicate survival less than a year, and positive (orange) features indicate survival greater than a year. The top 5 influential features for a specific test instance are shown. The weight of each feature (centre image) is used to calculate the prediction probability.

As we can see, the model classified the instance with 92% confidence. According to LIME, homogeneous morphology, younger age and a high KP Score were the most influential features for predicting survival less than a year. However, the features post-operative status and resection favored survival less than a year. One interpretation of this result is as follows: the patient may have a low grade, slow growing tumour which is unlikely to spread thus the condition is monitored rather than immediately treated [92]. With this type of approach, the patient is likely not at immediate risk of death, thus the model's prediction is logical. According to LIME, the random forest and logistic regression models had the same prediction probabilities, however the weighting of each feature varied slightly. The different feature weighting between models may explain their variation in performance (see Table 5.1).

5.2.4 Qualitative Analysis

The rule list models interpretability was assessed by Dr. Brennan and Mr. Poon. They were provided with multiple point-estimates from both BRL and FRL models (see Appendix B for all rule lists provided). To mitigate any potential bias, the models were constructed without the expert's input and only the final models were presented for evaluation. Note this was not an extensive evaluation, rather it was an informal evaluation where experts were asked to provide relatively high-level feedback about the rule lists.

In general, they both agreed that the rules produced were logical. The features used in the FRL, such as homogeneous morphology, high KP score, and meningioma, all favoured longer survival. Features such as glioblastoma, heterogenous morphology and older age, which favor survival less than a year, are not used by the FRL model. Instead, these features are summarised by the final rule, which predicts the probability of survival greater than a year to be 32%. On the other hand, the BRL contains features that favor both survival less than and greater than a year. As we can see, the BRL uses KP score less than 70, heterogenous morphology and brain metastases which indicates survival less than a year. Both Dr. Brennan and Mr. Poon agreed that the splitting of these feature types were in agreement with clinical knowledge.

5.2.5 Model Comparison

The features found important by the models for survival prediction are summarised visually by a Venn diagram in the Appendix (see Figure 6.8). Interestingly, all the

features used by the baseline Cox model were also found important by the rule list and ML models. The features identified as influential by all models included age, KP score, morphology, diagnosis, post-op status and first treatment. Both the rule lists and ML models found referral, resection and midline shift to be influential. Furthermore, only the rule lists found history of cancer, symptom 1, comorbidity and side significant, while only the ML models found the Scottish Index of Multiple Deprivation (SIMD) score, max tumour size and lobe important. Surprisingly, none of the models (including the sequential feature selector) found sex, symptom 1 duration or sign 1 to be influential for classification. Of these, the duration of the first symptom is most surprising. However, this feature is reported by patients and may prove unreliable as it is likely difficult to remember having a headache for 10 weeks versus 12 weeks, for example. The irrelevance of sex is in line with the literature which finds that although men are more likely to develop a brain tumour [60], that at least for glioblastomas, the likelihood of survival is relatively similar (6.8% for males and 8.3% for females) [93].

5.3 Rule Lists for Glioblastoma Prediction

The BRL and FRL algorithms were briefly investigated for the prediction of a glioblastoma. Note that all patients in the dataset have a brain tumour, hence the purpose of this classifier was to predict a glioblastoma against other tumour types. A point-estimate of the FRL is shown in Figure 5.7 (see Figure 6.13 in Appendix for BRL). Only patient demographics, medical history and symptom features were used for classification. The model had an accuracy of 0.692, an F1 score of 0.685 and an AUROC of 0.702.

IF KP Score: 90 AND Sex: Male THEN Glioblastoma risk is 58.84%, *Support*: 95
ELSE IF Symptom 1: Behavioral/Cognitive **AND** KP Score: 50-70 **THEN** Glioblastoma risk is 57.69%, *Support*: 52
ELSE IF Symptom 1: Behavioral/Cognitive **AND** PMH of cancer: No **THEN** Glioblastoma risk is 57.63%, *Support*: 59
ELSE IF Symptom 1 Duration: 0-7 weeks **AND** PMH of cancer: No **THEN** Glioblastoma risk is 47.81%, *Support*: 274
ELSE Glioblastoma risk is 16.0%, *Support*: 350

Figure 5.7: FRL example 1 for prediction of glioblastoma.

This was the first rule list to use a patients sex for classification. This supports current literature which finds that males are more at risk for glioblastomas than females [59]. Additionally, this was the first rule list to use symptom 1 duration as an informative feature. This may be due to the reduced number of features available to the model. Surprisingly, the first rule of the FRL, which classifies the most at-risk patients, finds the risk of a glioblastoma to be 59%. This is not a strong prediction and highlights the

difficulty of predicting brain tumour type based on presenting features alone.

Nonetheless, Dr. Brennan advised that this type of rule list may have significant clinical benefit but is limited by the current dataset. For example, a patient may present in primary care with non-specific symptoms. By the time the patient receives an official brain tumour diagnosis, often through a biopsy, the tumour may already be in an advanced state. This is a lengthy process. Thus the development of a non-invasive tool to predict the type of brain tumour may be of great utility. By reducing the time between diagnosis and start of treatment, a patient may have a better chance at survival.

5.4 Clinical Utility of Rule Lists

In terms of clinical utility, Dr. Brennan mentioned that the combination of features for both survival and glioblastoma prediction were informative and in-line with domain knowledge. However, although interesting on its own, he noted that the rules may not be of significant help given the current dataset. Dr. Brennan suggested that if this type of model was combined with other clinical information, such as blood tests, this may prove more powerful. Blood tests are now being investigated as a means for brain tumour diagnosis [94], and combined with the clinical information outlined by a rule list, the two may make a useful diagnostic tool.

Some of the advantages of rule lists in clinical practice are as follows: its innate interpretability, the simple *if...then...* structure of the rules is easy for clinicians to understand and predictions with rules are fast (only a few binary statements need to be assessed). However, these advantages are mitigated by the requirement of categorical features and the current use of rule lists for classification only instead of regression. Dr. Brennan also commented on the choice of the rules. For example, the first rule in the FRL for survival prediction is: *If* Morphology: Homogeneous *and* KP Score: 100 *then* probability of Survival $>$ 1 year is 97.37% (see Figure 5.2). Dr. Brennan mentioned that he would expect the second rule to be homogeneous tumour and KP score of 90. He regarded that the switching of different features "would not help with clinical decision making". The association rules are combined in a way that maximises the posterior (as discussed in Section 2.3.1). Thus the user does not have direct influence over the rule choice or order. One solution may be the incorporation of fewer features in the model. For example, we briefly explored the creation of a rule list to predict a glioblastoma for males only (and females only). The rule lists produced were similar, but given a larger dataset this approach may produce a more fine-grained set of rules.

Chapter 6

Conclusion

This project investigated the performance and interpretability of multiple algorithms for the prediction of brain tumour survival. We argued that interpretability is crucial for the implementation of ML algorithms in healthcare, and that a model's ability to explain its predictions is important for establishing a user's trust in the model. This was the first time classification algorithms were applied to the REDCap dataset to predict 1-year survival and glioblastoma diagnosis. Our results demonstrated that interpretable models, such as rule-lists, can perform on par with popular ML algorithms, while having the added benefit of interpretability.

The interpretability of multiple algorithms was assessed by evaluating the importance of dataset features for making predictions. We found that all six models used similar features, however the evaluation methods for feature importance varied between models. The three ML models were assessed using post-hoc interpretability methods which may be less reliable than intrinsically interpretable models. Post-hoc methods assess interpretability after model construction, thus explanations may be misleading or unreliable [11]. As we saw with random forests, depending on the evaluation method (e.g. MDI versus permutation importance), feature significance may be weighted differently thus producing inconsistent results. Additionally, there is a growing body of literature that has questioned the reliability of LIME and other post-hoc methods [11, 19, 95, 96, 97]. A model which is interpretable by design may provide more faithful explanations, however this approach is not without its own challenges. Innate interpretability is model-specific, which makes the comparison of interpretability between algorithms more difficult. In this project, the rule lists were also assessed qualitatively by domain experts. They provided high-level feedback on the rules produced and provided their opinion on the potential clinical utility of such a model. How-

ever access to experts is often limited and their time is precious, thus the significance of these results are typically restricted to a small number of observations. Nonetheless, there is a large toolbox of methods [98] to assess interpretability and this project has only scratched the surface.

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 proving trustworthy data-driven support. Interpretable models are a natural choice for the domain of predictive medicine, and integration of such models into clinical practice is an important first step in maximising patient survival.

6.1 Future Work

The design, development and integration of accurate interpretable ML models in healthcare serves as an important avenue for future research. Although our results are encouraging, a larger sample size is required for validation. The REDCap dataset is continuing to grow, and the addition of new data may allow the use of deep learning algorithms. An extension of this work could be the use of neural networks for survival prediction, assuming enough data becomes available. As neural networks are the quintessential black box algorithm, a comparative study of model-agnostic methods applied to neural networks or other ML algorithms would be an interesting next step.

As discussed in Section 4.1, the original goal of this project was to develop a rule list classifier for brain tumour diagnosis, but the CPRD dataset could not be secured in time. As such, in the future it would be beneficial to combine the CPRD and REDCap dataset to predict the presence of a brain tumour based on a patients presenting features. Dr. Brennan mentioned that this type of model may have greater clinical utility, especially for GPs, compared to a model predicting survival.

Rule lists have proven to be a powerful interpretable ML algorithm, however BRL and FRL are currently limited to binary classification. However, research into multiclass rule list algorithms is already underway [99]. The authors provided code was briefly explored in this project, but due to problems with hard-coding it was not investigated further. The prospect of multi-class predictions is an important area for future work. The creation of a rule list to predict the type of brain tumour based on a patients presenting features may be useful for triaging the urgency of patient referrals.

Appendix A

Glossary of Dataset Features

Age the age of a patient

Sex the sex of a patient

History of Cancer whether the patient has a past medical history of cancer

Comorbidity 1 the presence of another illness or disease occuring in a patient

Scottich Index of Multiple Deprivation (SIMD) a measure of deprivation of the area a patient lives

Karnofsky performance score (KP Score) a common measure in oncology to assess the functional state of a patient

Symptom 1 the first symptom type a patient presented with (reported by the patient)

Symptom 1 Duration the length of time of a patient's first symptom (minimum 0 weeks, maximum 52 weeks)

Symptom 2 the second symptom type a patient presented with (reported by the patient)

Sign 1 the first sign type a patient presented with (observed by the physician)

Urgency the patient's urgency of referral from primary care (emergency, suspicion of cancer, soon, routine)

Radiological Diagnosis (or Tumour Type) the type of brain tumour a patient was diagnosed with

Maximum Tumour Size a measure of the tumour size

Side the side of the brain the tumour is located (left, right, both, or midline)

Lobe the lobe where the tumour is located

Morphology the histological classification of the tumour based on the cell types present (homogenous, heterogenous)

Midline shift a measure of the brain's horizontal shift from the mid (centre) line

First Treatment the type of first cancer treatment (surgery, radiotherapy, chemotherapy)

Extent of Resection the amount of cancerous cells removed during surgery (measured in percentage)

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 6.6)

Supplementary Data Analysis

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 6.1: Sign domain classifications based on Dr. Brennan's recommendation. All examples are from the REDCap dataset.

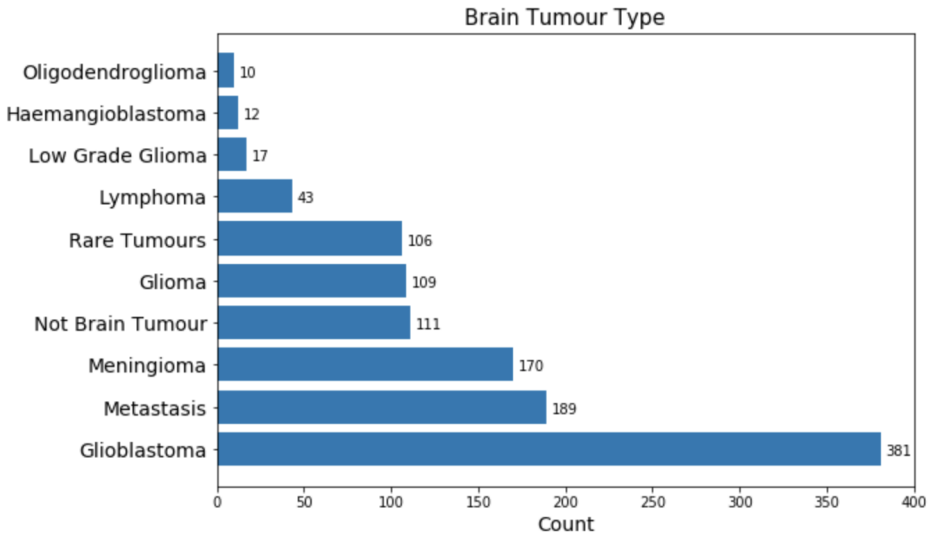

Figure 6.1: Proportion of patients with each type of cancer. The number of patients in each group is depicted beside the bar.

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.

Table 6.2: The original description of the Karnofsky performance status given by Karnofsky and Burchenal [74].

Target Variable	Mode-fill	Logistic Regression	KNN	KNN (norm)
SIMD Score	0.234	0.229	**0.246**	0.209
KP Score	0.383	**0.528**	0.456	0.517
Urgency of Referral	0.667	0.672	0.659	**0.683**
Side	0.424	**0.436**	0.422	0.406
Lobe	0.334	**0.360**	0.316	0.338
Morphology	0.676	**0.876**	0.769	0.826
Midline Shift	0.422	**0.434**	0.375	0.410
Extent of Resection	0.417	**0.482**	0.434	0.472
Post-op Status	0.473	**0.618**	0.568	0.573

Table 6.3: Accuracy results for categorical variable imputation.

Target Variable	Mean-fill	Linear Regression	KNN	KNN (norm)
Symptom 1 Duration	280.0	**237.9**	263.8	238.7
Max Tumour Size	311.7	258.6	294.6	**243.7**

Table 6.4: Mean square error results for continuous variable imputation.

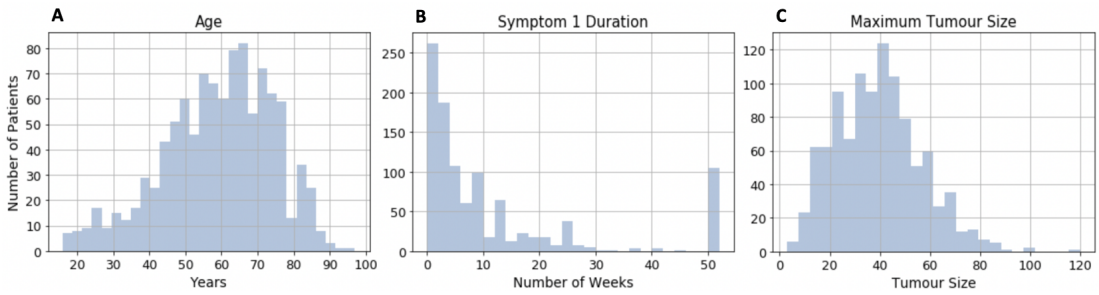

Figure 6.2: Distribution of features for discretisation.

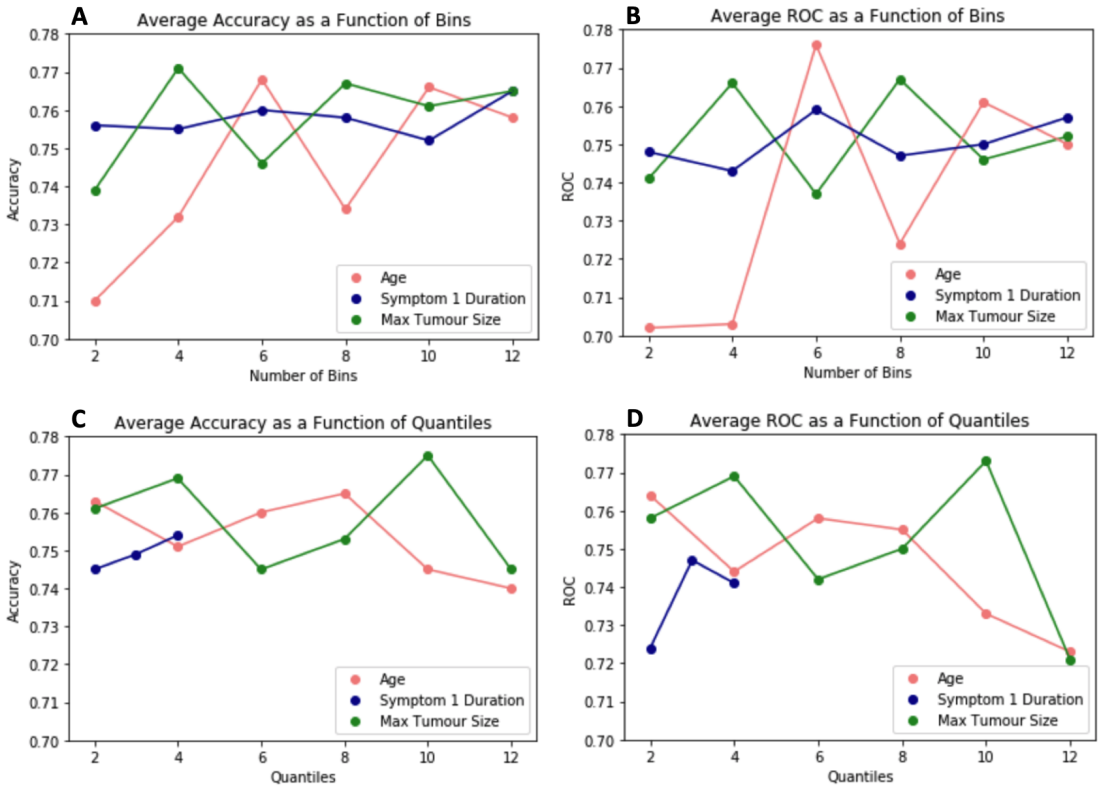

Figure 6.3: Average accuracy and area under ROC curve as a function of bins (A, B) and quantiles (C, D). Quantile discretisation for Symptom 1 Duration was only performed up to four quantiles due to the distribution of the feature.

Feature	Discretisation Method	Accuracy	F1 Score	ROC
Age	Bins: 6	**0.768**	**0.765**	**0.776**
	Quantile: 2	0.763	0.758	0.764
	Manual	0.789	0.793	0.796
Symptom 1 Duration	Bins: 6	**0.760**	**0.754**	**0.759**
	Quantile: 4	0.754	0.740	0.741
	Manual	0.759	0.746	0.747
Maximum Tumour Size	Bins: 4	0.771	0.763	0.766
	Quantile: 10	0.775	0.768	0.773
	Manual	**0.777**	**0.772**	**0.778**

Table 6.5: Results of continuous feature discretisation. The number of bins and quantiles that performed best was selected for each feature and compared to manual discretisation. The discretisation method that performed the best is highlighted in bold.

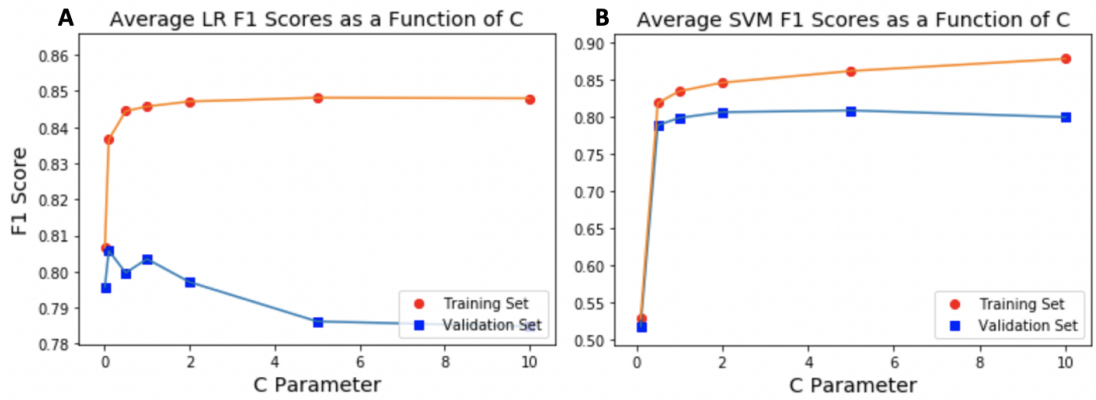

Figure 6.4: The affect of the regularisation parameter *C* on logistic regression (A) and SVM (B) performance. A value of 0.1 was selected for logistic regression and a value of 2 for SVM.

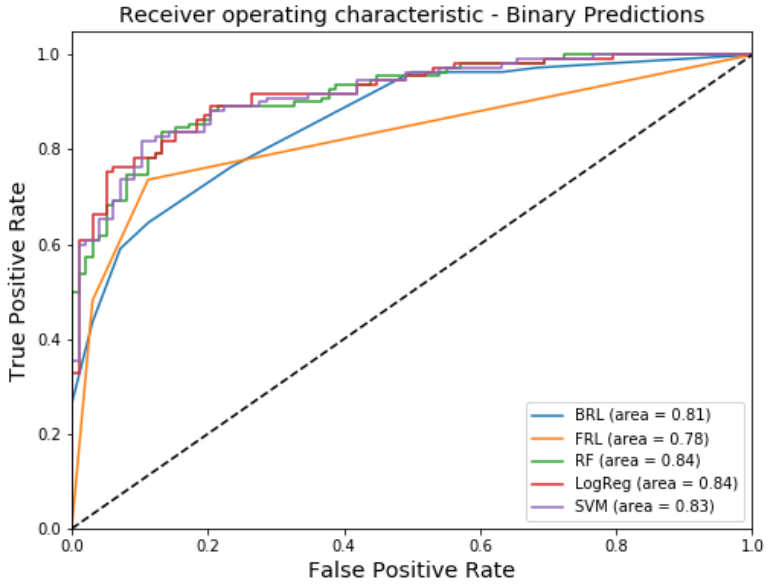

Figure 6.5: ROC curves for baseline, rule list and ML models evaluated on the test set. The area under the ROC curve was calculated using binary predictions.

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 6.6: Description of a patient's performance status (or functional state) developed by the Eastern Cooperative Oncology Group [100].

Rank	Feature
1	KP Score: 100
2	Symptom 1: Behavioral/Cognitive
3	Diagnosis: Glioblastoma
4	Diagnosis: Lymphoma
5	Diagnosis: Metastasis
6	Midline Shift: 0
7	Morphology: Heterogenous
8	Maximum Tumour Size: 0-18
9	Age: 50-65
10	Age: 65+
11	First Treatment: None
12	First Treatment: Radiotherapy
13	Post-op Status: 4
14	Post-op Status: No Surgery
15	Resection: 100%
16	Resection: No Resection

Table 6.7: Top 16 features returned using sequential feature selector.

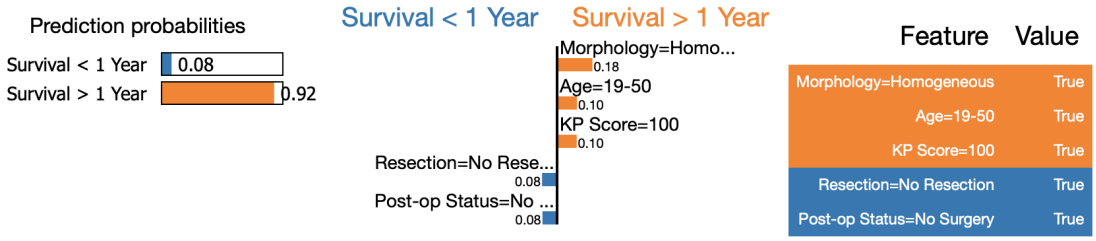

Figure 6.6: Feature importance of random forest model using LIME.

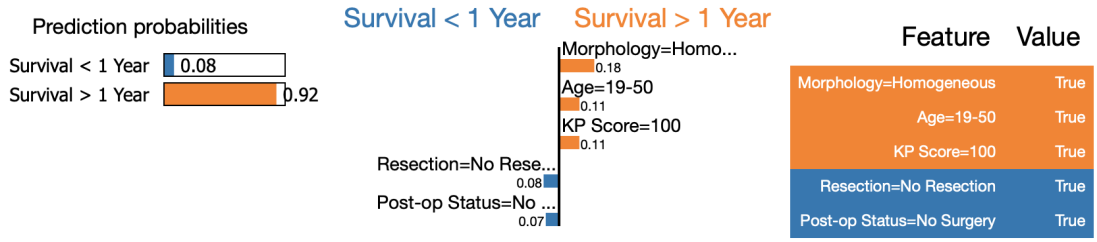

Figure 6.7: Feature importance of logsitic regression model using LIME.

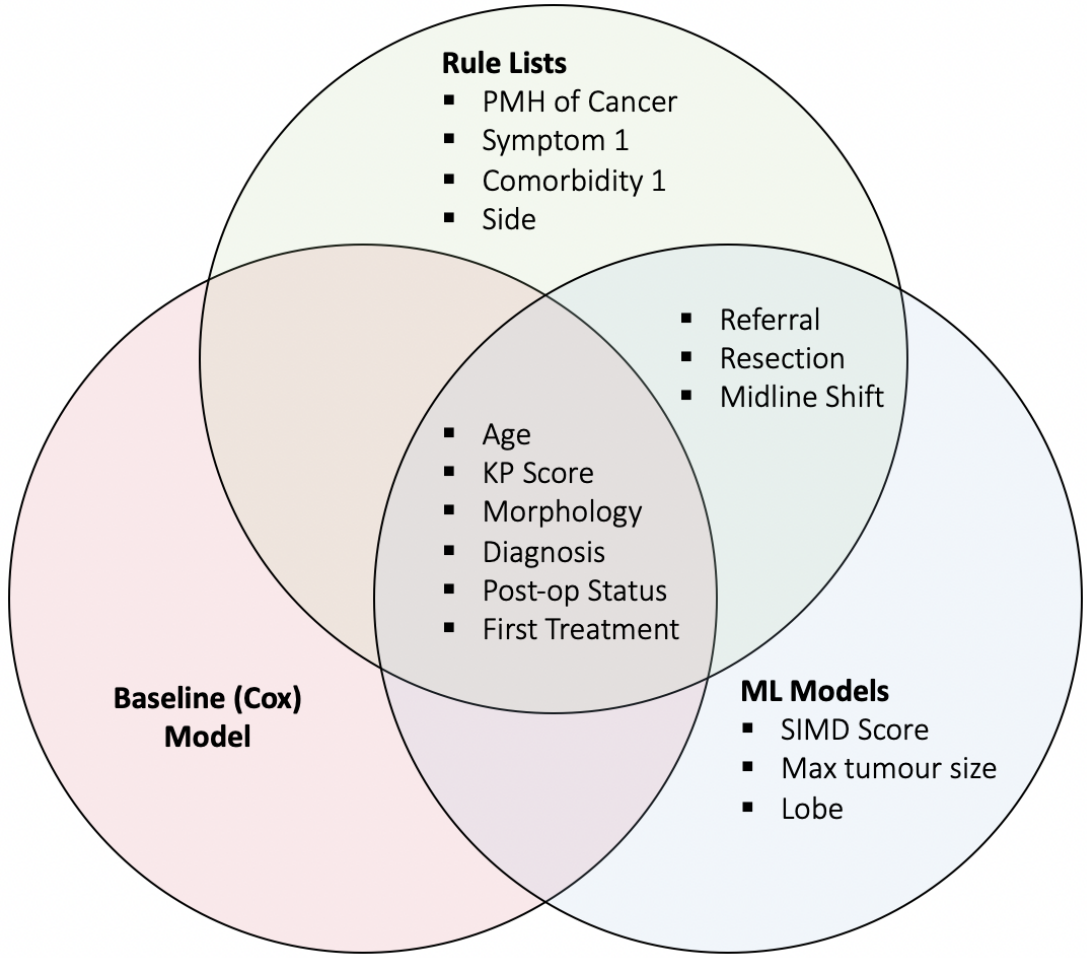

Figure 6.8: Venn diagram of features used by the three main groups of models for survival prediction. Features used by all model types are found in the centre.

Appendix B

IF Diagnosis : Metastasis **AND** Age : 65+ **THEN** probability of Survival > 1 year: 22.1% (13.6%-31.9%)
ELSE IF Post-op Status : 0.0 **AND** Morphology : Homogeneous **THEN** probability of Survival > 1 year: 97.4% (93.8%-99.5%)
ELSE IF First Treatment : None **AND** Diagnosis : Glioblastoma **THEN** probability of Survival > 1 year: 2.7% (0.3%-7.4%)
ELSE IF Diagnosis : Metastasis **AND** Urgency of referral : Emergency **THEN** probability of Survival > 1 year: 41.1% (28.7%-54.1%)
ELSE IF Resection : 100% **AND** First Treatment : Excision surgery **THEN** probability of Survival > 1 year: 96.8% (91.2%-99.6%)
ELSE IF Post-op Status : 0.0 **AND** Midline shift : 0 **AND** KP Score : 100 **THEN** probability of Survival > 1 year: 93.9% (83.8%-99.2%)
ELSE IF Diagnosis : Glioblastoma **AND** Resection : No Resection **AND** PMH of cancer : No THEN probability of Survival > 1 year: 22.1% (14.0%-31.4%)
ELSE IF Lobe : Frontal **AND** Age : 19-50 **THEN** probability of Survival > 1 year: 93.8% (85.5%-98.7%)
ELSE IF Morphology : Homogeneous **AND** Midline shift : 0 **AND** PMH of cancer : No **THEN** probability of Survival > 1 year: 94.1% (88.3%-98.0%)
ELSE probability of Survival > 1 year: 50.5% (44.7%-56.4%)

Figure 6.9: BRL example 1. Accuracy: 0.890. F1 Score: 0.883. AUROC: 0.876.

IF Diagnosis : Meningioma **AND** First Treatment : Excision surgery **THEN** probability of Survival > 1 year: 95.7% (90.9%-98.8%)
ELSE IF Age : 19-50 **AND** Urgency of referral : Emergency **THEN** probability of Survival > 1 year: 96.2% (89.7%-99.5%)
ELSE IF Resection : No Resection **AND** Diagnosis : Glioblastoma **THEN** probability of Survival > 1 year: 5.7% (1.6%-12.2%)
ELSE IF Symptom 1 : Behavioral/Cognitive **AND** KP Score : 50-70 **THEN** probability of Survival > 1 year: 9.8% (2.8%-20.4%)
ELSE IF Diagnosis : Metastasis **AND** Post-op Status : No Surgery **THEN** probability of Survival > 1 year: 18.8% (10.2%-29.1%)
ELSE IF Morphology : Homogeneous **AND** Midline Shift : 0 **THEN** probability of Survival > 1 year: 89.7% (83.7%-94.5%)
ELSE IF Midline Shift : 0 **AND** KP Score : 100 **THEN** probability of Survival > 1 year: 82.0% (70.3%-91.2%)
ELSE IF First Treatment : Biopsy **AND** Diagnosis : Glioblastoma **THEN** probability of Survival > 1 year: 23.0% (13.4%-34.2%)
ELSE IF Post-op Status : 0.0 **AND** Comorbidity 1 : NONE **THEN** probability of Survival > 1 year: 75.9% (66.2%-84.4%)
ELSE probability of Survival > 1 year: 47.9% (41.3%-54.6%)

Figure 6.10: BRL example 2. Accuracy: 0.783. F1 Score: 0.783. AUROC: 0.802.

IF Age : 19-50 **AND** Symptom 1 : Headache **THEN** probability of Survival > 1 year: 97.5% (91.0%-99.9%)
ELSE IF First Treatment : None **AND** Diagnosis : Glioblastoma **THEN** probability of Survival > 1 year: 1.5% (0.0%-5.6%)
ELSE IF Diagnosis : Metastasis **AND** PMH of cancer : No **THEN** probability of Survival > 1 year: 25.9% (15.3%-38.3%)
ELSE IF Resection : 100% **AND** PMH of cancer : No **THEN** probability of Survival > 1 year: 96.8% (92.5%-99.3%)
ELSE IF Side : Both Left and Right **AND** Symptom 1 : Behavioral/Cognitive **THEN** probability of Survival > 1 year: 15.4% (2.1%-38.5%)
ELSE IF KP Score : 50-70 **AND** Side : Right **THEN** probability of Survival > 1 year: 16.1% (5.6%-30.7%)
ELSE IF Morphology : Homogeneous **AND** PMH of cancer : No **THEN** probability of Survival > 1 year: 97.0% (93.6%-99.2%)
ELSE IF KP Score : 80 **AND** Symptom 1 : Headache **THEN** probability of Survival > 1 year: 10.0% (1.3%-26.0%)
ELSE IF Post-op Status : 0.0 **AND** Morphology : Heterogeneous **THEN** probability of Survival > 1 year: 72.0% (63.6%-79.7%)
ELSE IF First Treatment : Biopsy **AND** Diagnosis : Glioblastoma **THEN** probability of Survival > 1 year: 14.3% (5.6%-26.2%)
ELSE probability of Survival > 1 year: 47.1% (40.3%-53.9%)

Figure 6.11: BRL example 3. Accuracy: 0.770. F1 Score: 0.769. AUROC: 0.777.

IF Morphology: Homogeneous **AND** KP Score: 100 **THEN** probability of Survival > 1 year is 97.95%, *Support*: 146
ELSE IF Age: 19-50 **AND** PMH of cancer: No **THEN** probability of Survival > 1 year is 93.48%, *Support*: 46
ELSE IF Diagnosis: Meningioma **AND** PMH of cancer: No THEN probability of Survival > 1 year is 91.94%, *Support*: 62
ELSE IF Morphology: Homogeneous **AND** Midline Shift: 0 THEN probability of Survival > 1 year is 90.91%, *Support*: 22
ELSE IF First Treatment: Excision surgery **AND** KP Score: 100 THEN probability of Survival > 1 year is 82.86%, *Support*: 35
ELSE IF Post-op Status: 0.0 **AND** Morphology: Heterogeneous THEN probability of Survival > 1 year is 66.04%, *Support*: 106
ELSE probability of Survival > 1 year is 26.46%, *Support*: 378

Figure 6.12: FRL example 1. Accuracy: 0.780. F1 Score: 0.778. AUROC: 0.784.

IF Symptom 1 Duration : 46-52 weeks **AND** Sex : Female **THEN** probability of Glioblastoma: 3.5% (0.4%-9.6%)
ELSE IF Urgency of referral : routine (up to 12 weeks) **AND** KP Score : 100 THEN probability of Glioblastoma: 8.1% (1.8%-18.7%)
ELSE IF Age : 19-50 **AND** PMH of cancer : No **THEN** probability of Glioblastoma: 19.2% (11.1%-28.9%)
ELSE IF Symptom 2 : No Symptoms **AND** Sex : Female **THEN** probability of Glioblastoma: 20.7% (13.1%-29.4%)
ELSE IF Symptom 1 Duration : 7-13 weeks **AND** PMH of cancer : No **THEN** probability of Glioblastoma: 52.7% (43.4%-61.8%)
ELSE IF KP Score : 50-70 AND Symptom 1 : Behavioral/Cognitive **THEN** probability of Glioblastoma: 65.9% (51.5%-79.0%)
ELSE IF Symptom 1 Duration : 0-7 weeks AND PMH of cancer : No **THEN** probability of Glioblastoma: 59.5% (54.0%-64.9%)
ELSE IF KP Score : 50-70 AND Symptom 2 : Behavioral/Cognitive **THEN** probability of Glioblastoma: 55.6% (24.5%-84.3%)
ELSE probability of Glioblastoma: 16.7% (12.0%-21.9%)

Figure 6.13: BRL example 1 for prediction of glioblastoma. Accuracy: 0.692. F1 Score: 0.685. AUROC: 0.702.

Bibliography

[1] The Brain Tumour Charity. The statistics about brain tumours. 2020. Available at: https://www.thebraintumourcharity.org/get-involved/donate/why-choose-us/the-statistics-about-brain-tumours (Accessed: 6 June, 2020).

[2] American Cancer Society. Key Statistics for Brain and Spinal Cord Tumors. 2020. Available at: https://www.cancer.org/cancer/brain-spinal-cord-tumors-adults/about/key-statistics (Accessed: 6 June, 2020).

[3] Vincent KY Ho, Jaap C Reijneveld, Roelien H Enting, Henri P Bienfait, Pierre Robe, Brigitta G Baumert, Otto Visser, et al. Changing incidence and improved survival of gliomas. *European journal of cancer*, 50(13):2309–2318, 2014.

[4] Ahmad Faleh Tamimi and Malik Juweid. Epidemiology and outcome of glioblastoma. *Exon Publications*, pages 143–153, 2017.

[5] Richard Berk. *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media, 2012.

[6] Chih-Fong Tsai and Jhen-Wei Wu. Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert systems with applications*, 34(4):2639–2649, 2008.

[7] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5:8869–8879, 2017.

[8] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. *Jama*, 306(15):1688–1698, 2011.

[9] KC Johnston, AF Connors Jr, DP Wagner, WA Knaus, X-Q Wang, and E Clarke Haley Jr. A predictive risk model for outcomes of ischemic stroke. *Stroke*, 31(2):448–455, 2000.

[10] Jeffrey A Tice, Steven R Cummings, Rebecca Smith-Bindman, Laura Ichikawa, William E Barlow, and Karla Kerlikowske. Using clinical factors and mammographic breast density to estimate breast cancer risk: development and validation of a new predictive model. *Annals of internal medicine*, 148(5):337–347, 2008.

[11] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[12] Stefan Rüping. Learning interpretable models. Ph.D. thesis, Dortmund University of Technology. 2006.

[13] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

[14] Diogo V Carvalho, Eduardo M Pereira, and Jaime S Cardoso. Machine learning interpretability: A survey on methods and metrics. *Electronics*, 8(8):832, 2019.

[15] Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie 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*, page 1721–1730, New York, NY, USA, 2015. Association for Computing Machinery.

[16] Narges Razavian, Saul Blecker, Ann Marie Schmidt, Aaron Smith-McLallen, Somesh Nigam, and David Sontag. Population-level prediction of type 2 diabetes from claims data and analysis of risk factors. *Big Data*, 3(4):277–287, 2015.

[17] Cynthia Rudin and Berk Ustun. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces*, 48(5):449–466, 2018.

[18] Marco Tulio Ribeiro, Sameer Singh, and Carlos 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, pages 1135–1144, 2016.

[19] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. *arXiv preprint:1907.09294*, 2019.

[20] Edward Hance Shortliffe. Mycin: a rule-based computer program for advising physicians regarding antimicrobial therapy selection. Technical report, Stanford Univ Calif Dept of Computer Science, 1974.

[21] Gil Press. 12 AI Milestones: 4. MYCIN, An Expert System For Infectious Disease Therapy. *Forbes*, Apr 2020. Available at: https://www.forbes.com/sites/gilpress/2020/04/27/12-ai-milestones-4-mycin-an-expert-system-for-infectious-disease-therapy/ (Accessed: 13 July, 2020).

[22] David Earl Heckerman, Eric J Horvitz, and Bharat N Nathwani. Toward normative expert systems: Part i the pathfinder project. *Methods of information in medicine*, 31(02):90–105, 1992.

[23] Les Irwig, Petra Macaskill, Annabelle Farnsworth, R Gordon Wright, Jan McCool, Alexandra Barratt, and Judy M Simpson. A randomized crossover trial of papnet for primary cervical screening. *Journal of clinical epidemiology*, 57(1):75–81, 2004.

[24] Zhang Yue, Liang Fengchi, Su Fen, Bao Shounan, and Peng Yunxiang. A fuzzy production rule based expert system. *Fuzzy sets and systems*, 44(3):391–403, 1991.

[25] Robin Cowan. Expert systems: aspects of and limitations to the codifiability of knowledge. *Research Policy*, 30(9):1355–1372, 2001.

[26] Han Liu, Alexander Gegov, and Mihaela Cocea. *Rule based systems for big data: a machine learning approach*, volume 13. Springer, 2015.

[27] Benjamin Letham, Cynthia Rudin, Tyler H. McCormick, and David Madigan. Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model. *Annals of Applied Statistics*, 9(3):1350–1371, 2015.

[28] Fulton Wang and Cynthia Rudin. Falling rule lists. *Journal of Machine Learning Research*, 38:1013–1022, 2015.

[29] Robert C Holte. Very simple classification rules perform well on most commonly used datasets. *Machine learning*, 11(1):63–90, 1993.

[30] Jerzy Błaszczyński, Roman Słowiński, and Marcin Szelag. Sequential covering rule induction algorithm for variable consistency rough set approaches. *Information Sciences*, 181(5):987–1002, 2011.

[31] Johannes Fürnkranz, Dragan Gamberger, and Nada Lavrač. *Foundations of Rule Learning*. Springer Publishing Company, Incorporated, 2014.

[32] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine learning*, 29:131–163, 1997.

[33] Christian Borgelt. Frequent item set mining. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 2(6):437–456, 2012.

[34] Christian Borgelt. An implementation of the fp-growth algorithm. In *Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations*, pages 1–5, 2005.

[35] Mohammed Javeed Zaki. Scalable algorithms for association mining. *IEEE transactions on knowledge and data engineering*, 12(3):372–390, 2000.

[36] Rakesh Agarwal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, 1994.

[37] W Keith Hastings. *Monte Carlo sampling methods using Markov chains and their applications*. Oxford University Press, 1970.

[38] Peter JM Van Laarhoven and Emile HL Aarts. Simulated annealing. In *Simulated annealing: Theory and applications*, pages 7–15. Springer, 1987.

[39] Edney W Stacy et al. A generalization of the gamma distribution. *The Annals of mathematical statistics*, 33(3):1187–1192, 1962.

[40] Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216, 1994.

[41] David Roxbee Cox and David Oakes. *Analysis of survival data*, volume 21. CRC Press, 1984.

[42] Spotswood L Spruance, Julia E Reid, Michael Grace, and Matthew Samore. Hazard ratio in clinical trials. *Antimicrobial agents and chemotherapy*, 48(8):2787–2792, 2004.

[43] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[44] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[45] Leo Breiman. Some properties of splitting criteria. *Machine Learning*, 24(1):41–47, 1996.

[46] Scott Menard. *Applied logistic regression analysis*, volume 106. Sage, 2002.

[47] Magdalena Szumilas. Explaining odds ratios. *Journal of the Canadian academy of child and adolescent psychiatry*, 19(3):227, 2010.

[48] Johan AK Suykens and Joos Vandewalle. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300, 1999.

[49] Bernhard Schlkopf, Alexander J Smola, and Francis Bach. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. the MIT Press, 2018.

[50] David M Maslove, Tanya Podchiyska, and Henry J Lowe. Discretization of continuous features in clinical datasets. *Journal of the American Medical Informatics Association*, 20(3):544–553, 2013.

[51] Usama Fayyad and Keki Irani. Multi-interval discretization of continuous-valued attributes for classification learning. *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993.

[52] Ashley J Vargas and Curtis C Harris. Biomarker development in the precision medicine era: lung cancer as a case study. *Nature Reviews Cancer*, 16(8):525, 2016.

[53] Monica Arnedos, Cecile Vicier, Sherene Loi, Celine Lefebvre, Stefan Michiels, Herve Bonnefoi, and Fabrice Andre. Precision medicine for metastatic

breast cancer—limitations and solutions. *Nature Reviews Clinical Oncology*, 12(12):693, 2015.

[54] Siew-Kee Low, Hitoshi Zembutsu, and Yusuke Nakamura. Breast cancer: The translation of big genomic data to cancer precision medicine. *Cancer science*, 109(3):497–506, 2018.

[55] Su-In Lee, Safiye Celik, Benjamin A Logsdon, Scott M Lundberg, Timothy J Martins, Vivian G Oehler, Elihu H Estey, Chris P Miller, Sylvia Chien, Jin Dai, et al. A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia. *Nature communications*, 9(1):1–13, 2018.

[56] Alexia Iasonos, Deborah Schrag, Ganesh V Raj, and Katherine S Panageas. How to build and interpret a nomogram for cancer prognosis. *Journal of clinical oncology*, 26(8):1364–1370, 2008.

[57] Jill S Barnholtz-Sloan, Changhong Yu, Andrew E Sloan, Jaime Vengoechea, Meihua Wang, James J Dignam, Michael A Vogelbaum, Paul W Sperduto, Minesh P Mehta, Mitchell Machtay, et al. A nomogram for individualized estimation of survival among patients with brain metastasis. *Neuro-oncology*, 14(7):910–918, 2012.

[58] Haley Gittleman, Daniel Lim, Michael W Kattan, Arnab Chakravarti, Mark R Gilbert, Andrew B Lassman, Simon S Lo, Mitchell Machtay, Andrew E Sloan, Erik 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(5):669–677, 2017.

[59] Jigisha P Thakkar, Therese A Dolecek, Craig Horbinski, Quinn T Ostrom, Donita D Lightner, Jill S Barnholtz-Sloan, and John L Villano. Epidemiologic and molecular prognostic review of glioblastoma. *Cancer Epidemiology and Prevention Biomarkers*, 23(10):1985–1996, 2014.

[60] Haley Gittleman, Andrew E Sloan, and Jill S Barnholtz-Sloan. An independently validated survival nomogram for lower-grade glioma. *Neuro-Oncology*, 22(5):665–674, 10 2019.

[61] Olivier Graesslin, Bassam S Abdulkarim, Charles Coutant, Florence Huguet, Zsolt Gabos, Limin Hsu, Olivier Marpeau, Serge Uzan, Lajos Pusztai, Eric A

Strom, et al. Nomogram to predict subsequent brain metastasis in patients with metastatic breast cancer. *Journal of clinical oncology*, 28(12):2032–2037, 2010.

[62] Xiaowei Song, Arnold Mitnitski, Jafna Cox, and Kenneth Rockwood. Comparison of machine learning techniques with classical statistical models in predicting health outcomes. In *Medinfo*, pages 736–740, 2004.

[63] Ping Wang, Yan Li, and Chandan K. Reddy. Machine learning for survival analysis: A survey. *ACM Computing Surveys*, 51(6), 2019.

[64] Callum 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.

[65] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, and Carol Willing. Jupyter notebooks – a publishing format for reproducible computational workflows. In F. Loizides and B. Schmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press, 2016.

[66] Wes McKinney et al. Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX, 2010.

[67] Travis E Oliphant. *A guide to NumPy*, volume 1. Trelgol Publishing USA, 2006.

[68] John D Hunter. Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(3):90–95, 2007.

[69] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[70] Roger Stupp, Warren P Mason, Martin J Van Den Bent, Michael Weller, Barbara Fisher, Martin JB Taphoorn, Karl Belanger, Alba A Brandes, Christine Marosi,

Ulrich Bogdahn, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *New England Journal of Medicine*, 352(10):987–996, 2005.

[71] John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2006.

[72] Mio Ozawa, Paul M Brennan, Karolis Zienius, Kathreena M Kurian, William Hollingworth, David Weller, Robin Grant, Willie Hamilton, and Yoav 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(8), 2019.

[73] The Brain Tumour Charity. Adult brain tumour types. 2020. Available at: https://www.thebraintumourcharity.org/brain-tumour-diagnosis-treatment/types-of-brain-tumour-adult (Accessed: 1 August, 2020).

[74] David A Karnofsky, Walter H Abelmann, Lloyd F Craver, and Joseph H Burchenal. The use of the nitrogen mustards in the palliative treatment of carcinoma. with particular reference to bronchogenic carcinoma. *Cancer*, 1(4):634–656, 1948.

[75] AE Taylor, Ian N Olver, Thileepan Sivanthan, Marianne Chi, and Craig Purnell. Observer error in grading performance status in cancer patients. *Supportive Care in cancer*, 7(5):332–335, 1999.

[76] JB Sørensen, M Klee, T Palshof, and HH Hansen. Performance status assessment in cancer patients. an inter-observer variability study. *British journal of cancer*, 67(4):773–775, 1993.

[77] Lorenzo Beretta and Alessandro Santaniello. Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, 16(3):74, 2016.

[78] Amit Pandey and Achin Jain. Comparative analysis of knn algorithm using various normalization techniques. *International Journal of Computer Network and Information Security*, 9(11):36, 2017.

[79] Paul D Allison. Multiple imputation for missing data: A cautionary tale. *Sociological methods & research*, 28(3):301–309, 2000.

[80] Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M Ingersoll. An introduction to logistic regression analysis and reporting. *The journal of educational research*, 96(1):3–14, 2002.

[81] Judith E Dayhoff and James M DeLeo. Artificial neural networks: opening the black box. *Cancer: Interdisciplinary International Journal of the American Cancer Society*, 91(S8):1615–1635, 2001.

[82] G Ambler, S Seaman, and RZ Omar. An evaluation of penalised survival methods for developing prognostic models with rare events. *Statistics in medicine*, 31(11-12):1150–1161, 2012.

[83] Andrew Y Ng. Feature selection, l 1 vs. l 2 regularization, and rotational invariance. In *Proceedings of the twenty-first international conference on Machine learning*, page 78, 2004.

[84] Jooyoung Park and Irwin W Sandberg. Universal approximation using radial-basis-function networks. *Neural computation*, 3(2):246–257, 1991.

[85] Xinjian Guo, Yilong Yin, Cailing Dong, Gongping Yang, and Guangtong Zhou. On the class imbalance problem. In *2008 Fourth international conference on natural computation*, volume 4, pages 192–201. IEEE, 2008.

[86] Mohammad Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[87] Andrew P Bradley. The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, 30(7):1145–1159, 1997.

[88] Elizabeth B Claus, Kyle M Walsh, John K Wiencke, Annette M Molinaro, Joseph L Wiemels, Joellen M Schildkraut, Melissa L Bondy, Mitchel Berger, Robert Jenkins, and Margaret Wrensch. Survival and low-grade glioma: the emergence of genetic information. *Neurosurgical focus*, 38(1):E6, 2015.

[89] Eigil Husted Nielsen, Jörgen Lindholm, Peter Laurberg, Per Bjerre, Jens Sandahl Christiansen, Claus Hagen, Svend Juul, Jesper Jørgensen, Anders Kruse,

and Kirstine Stochholm. Nonfunctioning pituitary adenoma: incidence, causes of death and quality of life in relation to pituitary function. *Pituitary*, 10(1):67–73, 2007.

[90] Ibiayi Dagogo-Jack and Alice T Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology*, 15(2):81, 2018.

[91] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC bioinformatics*, 8(1):25, 2007.

[92] The Brain Tumour Charity. The statistics about brain tumours. 2020. Available at: https://www.thebraintumourcharity.org/brain-tumour-diagnosis-treatment/treating-brain-tumours/adult-treatments/watch-and-wait/ (Accessed: 7 August, 2020).

[93] Minjie Tian, Wenying Ma, Yueqiu Chen, Yue Yu, Donglin Zhu, Jingping Shi, and Yingdong Zhang. Impact of gender on the survival of patients with glioblastoma. *Bioscience reports*, 38(6), 2018.

[94] Ewan Gray, Holly J Butler, Ruth Board, Paul M Brennan, Anthony J Chalmers, Timothy Dawson, John Goodden, Willie Hamilton, Mark G Hegarty, Allan James, et al. Health economic evaluation of a serum-based blood test for brain tumour diagnosis: exploration of two clinical scenarios. *BMJ open*, 8(5), 2018.

[95] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 180–186, 2020.

[96] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *SafeAI@ AAAI*, pages 63–73, 2020.

[97] Eunjin Lee, David Braines, Mitchell Stiffler, Adam Hudler, and Daniel 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.

[98] Christoph Molnar. *Interpretable Machine Learning*. 2019. https://christophm.github.io/interpretable-ml-book/.

[99] Hugo M Proença and Matthijs van Leeuwen. Interpretable multiclass classification by mdl-based rule lists. *Information Sciences*, 512:1372–1393, 2020.

[100] Martin M Oken, Richard H Creech, Douglass C Tormey, John Horton, Thomas E Davis, Eleanor T McFadden, and Paul P Carbone. Toxicity and response criteria of the eastern cooperative oncology group. *American journal of clinical oncology*, 5(6):649–656, 1982.