Development of prediction models for one-year brain tumour survival using machine learning

Computer Methods and Programs in Biomedicine 233 (2023) 107482

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine

journal homepage: www.elsevier.com/locate/cmpb

Development of prediction models for one-year brain tumour survival using machine learning: a comparison of accuracy and interpretability

Colleen E. Charlton[a,*], Michael T.C. Poon[b,c,d,e], Paul M. Brennan[b,c,d], Jacques D. Fleuriot[a]

[a] Artificial Intelligence and its Applications Institute, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh EH8 9AB, UK
[b] Cancer Research UK Brain Tumour Centre of Excellence, CRUK Edinburgh Centre, University of Edinburgh, Edinburgh, UK
[c] Department of Clinical Neuroscience, Royal Infirmary of Edinburgh, 51 Little France Crescent EH16 4SA, UK.
[d] Translational Neurosurgery, Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK
[e] Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK

ARTICLE INFO

Article history:
Received 15 July 2022
Revised 15 December 2022
Accepted 12 March 2023

Keywords:
Bayesian rule lists
Explainable boosting machine
Interpretable models
Machine learning
Brain cancer
Survival

ABSTRACT

Background and Objective: Prediction of survival in patients diagnosed with a brain tumour is challenging because of heterogeneous tumour behaviours and treatment response. Advances in machine learning have led to the development of clinical prognostic models, but due to the lack of model interpretability, integration into clinical practice is almost non-existent. In this retrospective study, we compare five classification models with varying degrees of interpretability for the prediction of brain tumour survival greater than one year following diagnosis.

Methods: 1028 patients aged ≥16 years with a brain tumour diagnosis between April 2012 and April 2020 were included in our study. Three intrinsically interpretable 'glass box' classifiers (Bayesian Rule Lists [BRL], Explainable Boosting Machine [EBM], and Logistic Regression [LR]), and two 'black box' classifiers (Random Forest [RF] and Support Vector Machine [SVM]) were trained on electronic patients records for the prediction of one-year survival. All models were evaluated using balanced accuracy (BAC), F1-score, sensitivity, specificity, and receiver operating characteristics. Black box model interpretability and misclassified predictions were quantified using SHapley Additive exPlanations (SHAP) values and model feature importance was evaluated by clinical experts.

Results: The RF model achieved the highest BAC of 78.9%, closely followed by SVM (77.7%), LR (77.5%) and EBM (77.1%). Across all models, age, diagnosis (tumour type), functional features, and first treatment were top contributors to the prediction of one year survival. We used EBM and SHAP to explain model misclassifications and investigated the role of feature interactions in prognosis.

Conclusion: Interpretable models are a natural choice for the domain of predictive medicine. Intrinsically interpretable models, such as EBMs, may provide an advantage over traditional clinical assessment of brain tumour prognosis by weighting potential risk factors and their interactions that may be unknown to clinicians. An agreement between model predictions and clinical knowledge is essential for establishing trust in the models decision making process, as well as trust that the model will make accurate predictions when applied to new data.

1. Introduction

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

At the group level, basic statistical models are well-established in brain tumour survival analysis, but patient level survival prediction remains a challenge, possibly because of the well-known biological heterogeneity of the disease and of treatment responses. Machine learning (ML) approaches have recently demonstrated their utility in brain tumour survival analyses [3] (see also Kourou et al. [4] a general review), but such studies often use genetic or imaging data, with complex black box models to make prognostic

* Corresponding author.
E-mail address: colleen.charlton@camh.ca (C.E. Charlton).

https://doi.org/10.1016/j.cmpb.2023.107482

predictions. In clinical practice, though, this type of data is infrequent, hence these models are of limited applicability. Fulop et al. [5] used a large clinical and molecular brain tumour dataset to predict 400-, 900- and greater than 900-day survival after surgery, finding that a neural network achieved 63% accuracy, outperforming several other ML methods. Using the interpretability technique LIME (Local Interpretable Model-Agnostic Explanations) [6], the authors also investigated the main drivers behind miss-classified predictions and emphasized the importance of model interpretability for clinical decision-making. Senders et al. [7] recently used demographic, socioeconomic, clinical, and radiographic features for the creation of an online calculator for the prediction of glioblastoma survival. The authors compared 15 ML and statistical algorithms and an Accelerated Failure Time [8] algorithm was selected. However, a follow-up Letter to the Editor in the same journal noted inconsistencies in the model calculations and highlighted the danger of non-healthcare professionals using this online resource [9].

In data-driven healthcare, there is a desire for interpretable and explainable AI/ML models to give end users (e.g. clinicians) the support that will allow them to make informed judgements when it comes to high-stake medical decisions (see Ahmad et al. [10] for a review). However, there is no all-purpose definition of interpretability since this is a subjective concept that is often domain-specific [11]. One way to define interpretability is as the degree to which a human can understand the cause of a decision [12,13]. Interpretability may be achieved by either using an intrinsically interpretable model whereby its simple structure allows end users to understand feature relationships and final predictions, or by applying post-hoc explanation techniques to analyse and extract information from a trained model [14].

In this paper we investigate and compare the performance of three intrinsically interpretable 'glass box' classifiers, Bayesian Rule Lists (BRL) [15], Explainable Boosting Machines (EBM) [16] and Logistic Regression (LR) [17], and two 'black box' classifiers, Random Forest (RF) [18] and Support Vector Machine (SVM) [19], for the prediction of one year survival following a brain tumour diagnosis. We used SHapley Additive exPlanations (SHAP) values [20] to infer global feature importance of black box models and to understand misclassified predictions. Furthermore, clinical experts reviewed model feature importance and evaluated clinical utility.

2. Methods

The following section describes the raw dataset and its preprocessing, and the modelling steps for the prediction of brain tumour survival greater than one year. Brief explanations of the employed machine learning algorithms and interpretability techniques are also provided. We adhered to the TRIPOD reporting guideline for prediction models (see supplementary materials) [21].

Data was analysed in Python 3.8.1 [22]. Standard data analysis libraries, namely Pandas [23], NumPy [24], Matplotlib [25] and Scikit-learn [26] were used, as well as SHAP [20] for global feature importance. InterpretML [16], an open-source Python package for state-of-the-art ML interpretability techniques, was used for EBM and local-level SHAP implementation. LIME is another popular post-hoc interpretability tool [6], however due to the instability of the explanations [27], and the superior robustness of SHAP [28], this method was not investigated further. The BRL model was the original authors' code [15][1].

2.1. The raw dataset

In this retrospective study, we investigated an anonymised dataset collected from electronic patient records of consecutive patients identified by regional neuro-oncology multidisciplinary teams in Scotland, UK, between 1 April 2012 and 30 April 2020. Data collection was approved by Southeast Scotland Research Ethics Committee (reference 17/SS/0019).

The raw dataset contained 1283 patient records and 225 predictor variables. All patients were $\geq$ 16 years of age with a brain tumour diagnosis. Patients were excluded if they had incomplete records (n = 36) or lacked symptomatology information (n = 219). Of the 225 predictor variables, the majority was sparse, and pertained to specific symptom details (e.g. Symptom 1 - Headache [choice=Worse on coughing/bending]; n = 107), blood test and treatment particulars (e.g. baseline haemoglobin, date of chemotherapy start; n = 75) and other irrelevant information (e.g., location of first imaging; n = 23) (see Appendix *A.1: Data Preprocessing* for additional details). These variables were removed from the dataset. Based on exclusion criteria, the dataset was reduced to 1028 patients records, 20 predictor variables and one dependent variable, namely patient survival in days.

Patient survival was measured in days from radiological diagnosis of a brain tumour and in our dataset, 35% of patients were still alive (i.e. no death date is recorded). All predictive models (see Section 2.4) performed binary classification of survival greater than one year since this is a medically relevant timepoint: only around 40% of patients with a brain tumour survive beyond this. One year survival labels were thus created from the dependent variable i.e., patient survival in days. Our resulting dataset was (as expected) relatively even split: 443 patients (43%) survived less than a year and 585 patients (57%) survived more than a year.

2.2. Pre-processing

Following the removal of irrelevant variables, various preprocessing steps were needed to get the narrowed-down dataset of 1028 patient records into a state suitable for our analyses. To prevent data leakage, all preprocessing was embedded within the nested cross-validation (CV) procedure (see Section 2.3), hence all data preprocessing methods were prepared on the training set and applied to the train and test sets.

Given the small dataset size, imputation rather than deletion was employed to manage missing data. 11 of the 20 features had missing data, with a mean missingness of 7% $\pm$ 4%. All missing data was believed to be missing at random (see Appendix *A.1.1:Missing Data*). Multivariate imputation was implemented through the iterative imputer in Scikit-learn [26], which performs multiple imputation rounds whereby each feature with missing values is modelled as a function of other features and the final estimate is used for imputation. The iterative imputer cannot handle categorical variables, thus missing nominal variables (n = 2) were first imputed using the most frequent value. Missing ordinal variables (n = 7) were encoded as integer values and a Bayesian ridge regression estimator [29] with 10 imputation rounds was used for multivariate imputation. Missing values were initialized using the feature mean and the feature with the fewest missing values was imputed first. This method is similar to the R MICE package (Multivariate Imputation by Chained Equations) [30], however the iterative imputer differs by returning a single imputation instead of multiple imputations.

Following imputation, we tested for statistically correlated variables. Two sets of features, *Tumour Type* and *Likely Grade,* as well as *First Treatment* and *Extent of Resection,* were combined into a single feature called *Diagnosis* and *First Treatment*, respectively, to reduce feature collinearity (see Appendix *A.1.2: Feature Correlations* for more details). Finally, in order to support the association rule mining employed by the BRL algorithm (see Section 2.4.1), all continuous variables (*Age, Symptom 1 Duration*, and *Maximum Tumour Size*) were discretised. Most rule-mining approaches make the (re-

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

strictive) assumption that all features are binary or categorical, although some approaches automatically attempt to discretise continuous features [31]. Automatic discretisation of the continuous features was attempted, however the results did not agree with expert knowledge and literature, hence continuous features were discretised based on meaningful manually defined cut-points (see Appendix *A.1.3: Feature Discretization*). The other ML algorithms considered in this study (RF, LR, SVM and EBM) do not make this assumption, so the original continuous features were used in these cases. All continuous and ordinal variables were normalized into the range [0,1].

The final dataset contained 18 predictor variables including patient demographics (e.g., sex, age), medical history (e.g., history of cancer, comorbidity, Karnofsky performance score (KPS) - a measure of a patient's general well-being [32]), symptom features (e.g., symptom types and duration), radiological tumour analysis (e.g., diagnosis, maximum tumour size) and treatment details (e.g., first treatment, post-op performance status). A detailed overview of each feature, including their descriptions, values and proportion, is provided in Appendix *A.2: Description of Dataset Variables*.

2.3. Modelling pipeline

Since the EBM, LR, RF and SVM models cannot directly handle categorical variables (a technical constraint imposed by scikit-learn toolkit), a combination of binary (n=3), ordinal (n=5) and one-hot encoding (n=7) was used for the 15 categorical features, which resulted in a total of 58 different feature types. Due to the small dataset size, nested CV [33] was implemented for preprocessing, hyperparameter tuning and model assessment (see Appendix *A.3: Hyperparameter Tuning* for description of hyperparameter searches). The outer loop, which was responsible for assessing model performance, used 5-fold stratified CV resulting in an 80% training set (n = 822) and 20% test set (n = 206). The inner loop, which was responsible for hyperparameter tuning and model selection, used 3-fold stratified CV, resulting in a further split of the outer training set, into a 66% train set (n = 543) and 33% validation set (n = 279). Stratified CV ensures the training and test data in each fold reflect the distribution of the outcome variable (i.e., 1-Year Survival) from the dataset. Nested CV was repeated for three different random seeds resulting in a 3 × 5 × 3 setup. For each model, the average balanced accuracy (BAC), macro-F1 [34], sensitivity, specificity and area under the receiver operating characteristic curve (AUROC) [35] were reported.

2.4. Machine learning

2.4.1. Bayesian rule lists

BRLs are a type of rule list classification model that produce a series of *if-then* rules, also known as decision (or production) rules, where the goal is to learn $P(Y = 1|X)$. Y is binary, and in this analysis, $Y = 1$ would indicate survival greater than one year and X would represent a patient's features. The conditional probability distribution is represented as a decision list consisting of a series of decision rules.

The creation of a BRL roughly follows these steps: first, antecedents are extracted from the data using the frequent itemset mining technique FP-Growth [36] and second, a set of rules and their order are learned using Bayesian statistics. BRLs create a posterior distribution over rule lists, given the observed data and user specified priors, which are often used to favour concise rule lists with a small number of conditions. Using a generative model, an initial decision list is selected and iteratively modified using Markov chain Monte Carlo sampling [37] to generate many samples of decision lists from the posterior distribution. For each rule, 95% credible intervals are estimated using the Dirichlet distribution function [38] (see Letham et al., [15] for technical details). This procedure ensures the production of a variety of lists that are not dependent on one initial decision list. Given this posterior distribution of decision lists, new observations are classified using a point estimate (a single decision list) or the posterior predictive distribution (multiple decision lists). The point estimate is chosen as the list with the highest posterior probability from all the samples with posterior mean list length and posterior mean average rule cardinality.

2.4.2. Explainable boosting machine

EBM is a glass box model designed to have accuracy on-par with state-of-the-art ML methods while remaining highly explainable [16]. EBMs are a type of Generalised Additive Model (GAM) [39] with automatic interaction detection. GAMs model the impact of each predictive feature through smooth feature functions which, depending on the underlying data pattern, can be linear or nonlinear. However, GAMs use each feature individually, missing any relation between two features. EBMs improve upon traditional GAMs by utilizing modern ML techniques, namely bagging and gradient boosting, to learn the best feature function, and automatically detect and include pairwise feature interactions through round-robin cycles (such models are called GA^2M: Generalized Models with Interactions).

2.4.3. Other ML algorithms

Finally, the following three popular ML algorithms were utilized for brain tumour survival prediction: a LR classifier [17], a RF classifier [18] and an SVM classifier [19] with a radial basis function kernel [40]. All models were implemented using the scikit-learn package [26].

2.5. Model interpretability

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 making model comparison challenging. In medicine, where human decision making governs the process of patient treatment, a survey over domain experts is a valuable measure of interpretability. The interpretability of all our models was evaluated based on the clinical expertise of two of the authors (M.P. and P.B.). To mitigate any potential bias, the models were constructed without the expert's input and only the final models were presented for qualitative feedback. For the interpretability analysis, all ML models were trained on the same dataset (i.e. train/validation/test split) to ensure interpretability methods were consistently evaluated on the same training examples and local predictions could be fairly compared. Specifically, model interpretability was evaluated in the following ways:

- **BRL:** The experts were given five BRL point-estimates Appendix Bfor review and asked to consider whether: i) the rules produced were sensible, ii) any rules were surprising or unrealistic, and iii) the potential employability of such a model in a clinical setting.
- **EBM:** These models are highly explainable because the contribution of each feature to a final prediction can be visualized and understood by plotting the feature function (see Section 2.4.2). To understand individual predictions, each feature function serves as a lookup table per feature and returns a contribution term which can be sorted and visualized to understand individual feature importance. Features with larger positive or negative scores have a greater effect on the resulting prediction than features that have scores closer to zero.

Table 1
Performance metrics were assessed using nested cross-validation for three different random seeds. Mean and, in parenthesis, the standard deviation (SD) for 15 models are given. BRL: Bayesian Rule List, EBM: Explainable Boosting Machine, LR: Logistic Regression, RF: Random Forest, SVM: Support Vector Machine, AUROC: area under the receiver operating characteristic curve.

	BRL (SD)	EBM (SD)	LR (SD)	RF (SD)	SVM (SD)
Balanced Accuracy	0.726 (0.041)	0.771 (0.041)	0.775 (0.045)	0.789 (0.033)	0.777 (0.047)
Macro-F1	0.718 (0.040)	0.770 (0.040)	0.772 (0.042)	0.790 (0.033)	0.773 (0.049)
AUROC	0.780 (0.031)	0.864 (0.034)	0.867 (0.032)	0.878 (0.022)	0.865 (0.037)
Sensitivity	0.706 (0.078)	0.811 (0.046)	0.810 (0.067)	0.844 (0.036)	0.806 (0.078)
Specificity	0.746 (0.130)	0.731 (0.107)	0.740 (0.137)	0.734 (0.081)	0.748 (0.136)

RF, LR and SVM: SHAP, a post-hoc explainability tool, was used to assess global feature importance and individual predictions [20]. SHAP is a model-agnostic method that calculates the contribution of each feature using game theoretically optimal Shapley values [41], which are the average marginal contribution of a feature across all possible permutations. In other words, Shapley values considers all possible feature combinations for all possible model predictions. Due to this exhaustive search, SHAP can be computationally expensive, but provides theoretical guarantees for the consistency and accuracy of explanations, often making SHAP a preferred post-hoc explainability method [14].

Note that LR interpretability can be assessed using the odds ratio [42], or by taking the exponent of the model's learned feature weights. However, the interpretation of these weights is multiplicative and dependent on the feature type (e.g. numerical, binary categorical, categorical with more than two categories), making model interpretation challenging. For ease of model comparison, we report LR feature importance based on SHAP values only.

3. Results

3.1. Model evaluation

The mean classification performance of the five modelling approaches on the brain tumour dataset are summarised in Table 1. All models performed above the no-information rate of 56.9%, or the accuracy achieved by always predicting the majority outcome label. RF narrowly performed the best, with a BAC of 78.9%, closely followed by SVM (77.7% BAC), LR (77.5% BAC) and EBM (77.1% BAC). BRL performed slightly worse with a BAC of 72.6%.

3.2. Model interpretability

We compared interpretability across all models at a global model level and if possible, a local prediction level, going from the most-interpretable glass box models to the least-interpretable black box models.

3.2.1. Bayesian rule list interpretability

Fig. 1 shows a BRL point-estimate that obtained a BAC of 78.2%, the highest from the first random seed. For a given rule list, once a patient has satisfied a rule they will not be taken into account by the cases further down the list. The final rule in the list will only consider the subset of patients that were not classified by the previous ones. The BRL point-estimates from the four other folds are given in Appendix B, for a total of five BRL estimates that were presented to the experts. Despite the different point-estimates, there is significant overlap in the rules. Due to the iterative construction of BRLs, multiple equally good rule lists may be produced and it is not determinable which will be returned by the model ahead of time [15].

The given BRL point-estimate (Fig. 1) indicates that if a patient with a brain tumour has a KPS <= 70 and tumour on the right side of their brain, the probability they will survive more than one year is 25.5%. However, if this is not true, and the patient is diagnosed with a benign meningioma, the probability they will survive more than one year is 98.9%. Otherwise, if the patient is diagnosed with a benign glioma, this is indicative of survival greater than one year with a probability of 98.3%.

According to the qualitative evaluation by the experts, there is no gold standard of expected feature rankings, but the combination of features used by the rule lists for survival prediction are informative and in-line with domain knowledge. The rule lists re-produced feature combinations that are well-established in the literature [43,44,45], and although the rule lists did not uncover novel feature relationships, an agreement between model predictions and clinical knowledge is essential for establishing trust in the models decision making process, as well as trust that the model will make accurate predictions when applied to new data. Furthermore, the rule lists simple IF-THEN structure was easy to interpret and classification of a patient likely to be fast, since only a few statements need to be reviewed. However, each statement switches between feature types, and the probability of survival does not follow a linear order, making the interpretation more cumbersome. Finally, the integration of additional clinical information, such as blood tests or genetic data, was suggested to improve the clinical validity of rule list models. Blood tests are now being investigated as a means for brain tumour diagnosis [46] and genetic alterations have shown to be effective predictors for tumour prognosis [5,45].

3.2.2. Global model interpretability

The interpretability of the remaining four models, EBM, LR, RF and SVM, was evaluated using global explainability methods, through which the average behaviour of the model is described. EBM's absolute global feature importance was assessed using the model's learned feature function and the global feature importance for LR, RF and SVM was assessed using SHAP. Fig. 2 depicts the top 12 informative features used by the respective models. In general, all models found age, diagnosis, functional features (Karnofsky Performance Status [KPS] and post-op performance status) and first treatment to be the most influential. Other informative features included comorbidity, Scottish Index of Multiple Deprivation (SIMD), Symptom 1 duration and tumour side, although this varied between models. Fig. 3 takes this a step further and visualises the impact of different feature values on the model's output. For example, across all models, a younger age (blue) has a positive impact on survival, or higher post-op performance scores (red) - i.e., worse functional state following surgery - have a negative impact on survival. In Fig. 3, for model comparison, EBM feature importance was also evaluated with SHAP.

3.2.3. Local model interpretability

EBM, LR, RF and SVM models were also evaluated using local explainability methods, which are beneficial for understanding in-

	Condition		Probability	Credible Interval	Support
IF	KPS <=70 **AND** Side: Right	THEN chance of survival > 1 Year	25.5%	(15.0%-37.6%)	55
ELSE IF	Diagnosis: Meningioma Benign	THEN chance of survival > 1 Year	98.9%	(96.0%-100.0%)	92
ELSE IF	Diagnosis: Glioma Benign	THEN chance of survival > 1 Year	98.3%	(93.7%-100.0%)	59
ELSE IF	Age < 45	THEN chance of survival > 1 Year	91.7%	(84.9%-96.5%)	83
ELSE IF	First Treatment: Surgery Removal 90-99%	THEN chance of survival > 1 Year	63.2%	(54.1%-71.7%)	120
ELSE IF	Diagnosis: Metastasis **AND** Urgency of Referral: Emergency	THEN chance of survival > 1 Year	18.6%	(9.9%-29.4%)	55
ELSE IF	First Treatment: Surgery Removal 100%	THEN chance of survival > 1 Year	90.2%	(76.6%-97.2%)	40
ELSE IF	Maximum Tumour Size < 20mm **AND** Midline Shift: 0	THEN chance of survival > 1 Year	68.6%	(52.5%-82.6%)	30
ELSE IF	Age 60-74 **AND** Comorbidity: Yes	THEN chance of survival > 1 Year	14.5%	(7.8%-22.7%)	83
ELSE IF	KPS: 100	THEN chance of survival > 1 Year	63.0%	(48.8%-76.2%)	43
ELSE IF	First Treatment: None	THEN chance of survival > 1 Year	6.1%	(1.7%-12.9%)	62
ELSE		THEN chance of survival > 1 Year	46.0%	(37.0%-55.2%)	100

Fig. 1. BRL-point estimate with the highest balanced accuracy of 78.2% obtained from the first random seed. The credible interval refers to the 95% probability that given the data the outcome variable (survival > 1 year) falls within the interval. Support refers to the number of patient records supporting the given rule. KPS: Karnofsky Performance Score.

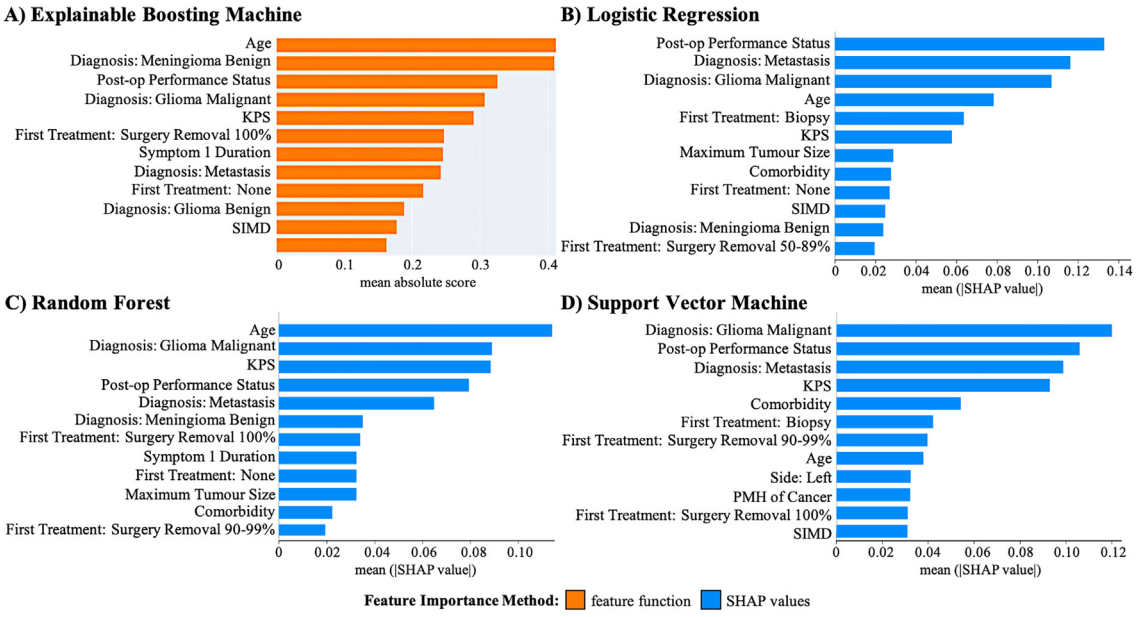

Fig. 2. Top 12 influential features for each model. (A) Explainable Boosting Machine's absolute global feature importance is based on feature scores determined by individual feature functions. Global feature importance for (B) Logistic Regression (C) Random Forest and (D) Support Vector Machine was based on the mean absolute SHAP (SHapley Additive exPlanation) value. All features are described in Appendix A.2: Description of Dataset Variables. KPS: Karnofsky Performance Score, SIMD: Scottish Index of Multiple Deprivation, PMH: Previous medical history.

dividual predictions, especially misclassified predictions. EBM local feature importance was assessed using the model's learned feature function and LR, RF and SVM local feature importance was assessed using SHAP. In the interest of brevity, only one patient record is discussed here, and an additional example is provided in Appendix *A.4.1: Local Feature Importance*. Fig. 4**.** illustrates the 10 most influential features for the classification of a single patient. For the given patient, EBM, LR and SVM correctly classified the patient, and RF did not. One-hot encoded data was used by the models to make predictions thus both the presence (value = 1) and absence (value = 0) of a feature is used to make a prediction.

Across all models, influential features found by the global interpretability methods, such as diagnosis, first treatment and functional status, were also informative for local predictions. Additional informative features for the given patient include age, symptom information and the presence of a comorbidity. We can investigate RF's misclassification by evaluating SHAP's explanation. One interpretation of this result is as follows: the patient was older (age 75) with a malignant glioma and able to carry on normal activities with effort (KPS 80), all of which had a negative impact on survival. However, the patient had no comorbidities, underwent a tumour resection of 90-99%, and was capable of light work and activity (post-op status of 1) following surgery, conceivably leading to survival greater than one year. By referring back to Fig. 2C and Fig. 3C, we know that RF places the greatest importance on age, diagnosis and KPS, potentially leading to the overweighing of these negative features for the given patient.

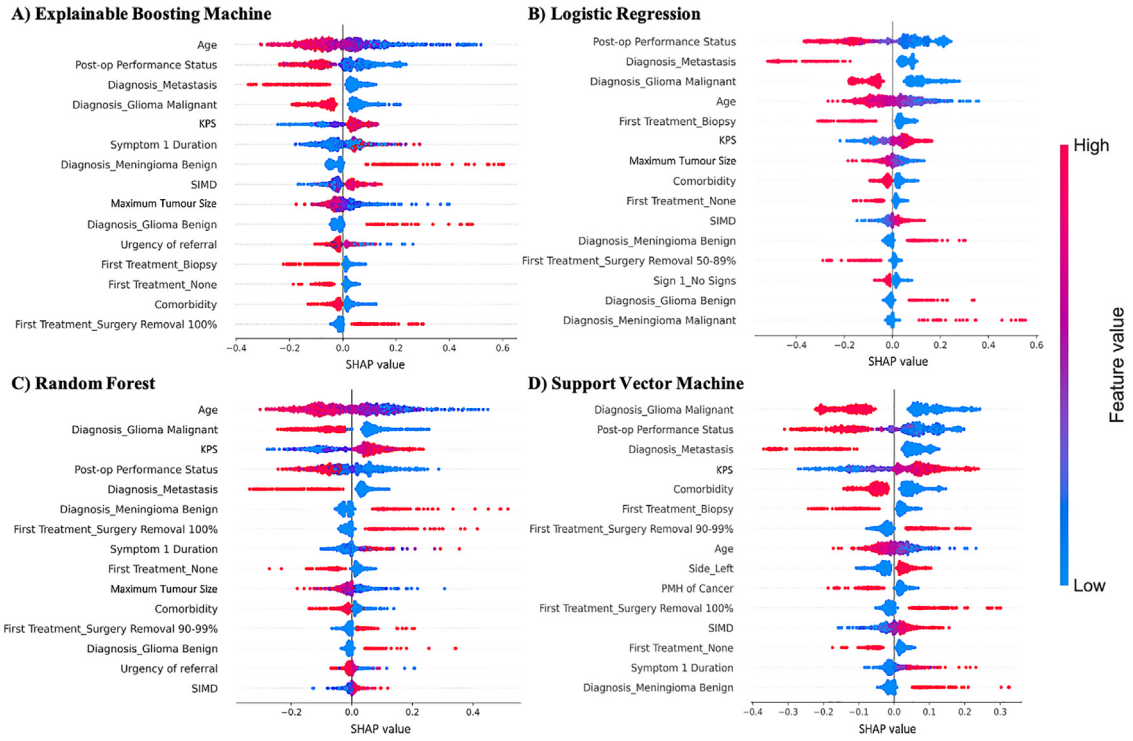

Fig. 3. SHAP (SHapley Additive exPlanation) value summary plots of the top 15 most influential features for each model. For each feature, one dot corresponds to a single patient. The dot's position along the x-axis represents the impact the feature had on the model's output for that specific patient. A patient with a higher SHAP value has a higher change of survival greater than one year, compared to a patient with a lower SHAP value. Features are ordered along the y-axis based on their average absolute importance (see Fig. 2). KPS: Karnofsky Performance Score, SIMD: Scottish Index of Multiple Deprivation, PMH: Previous medical history.

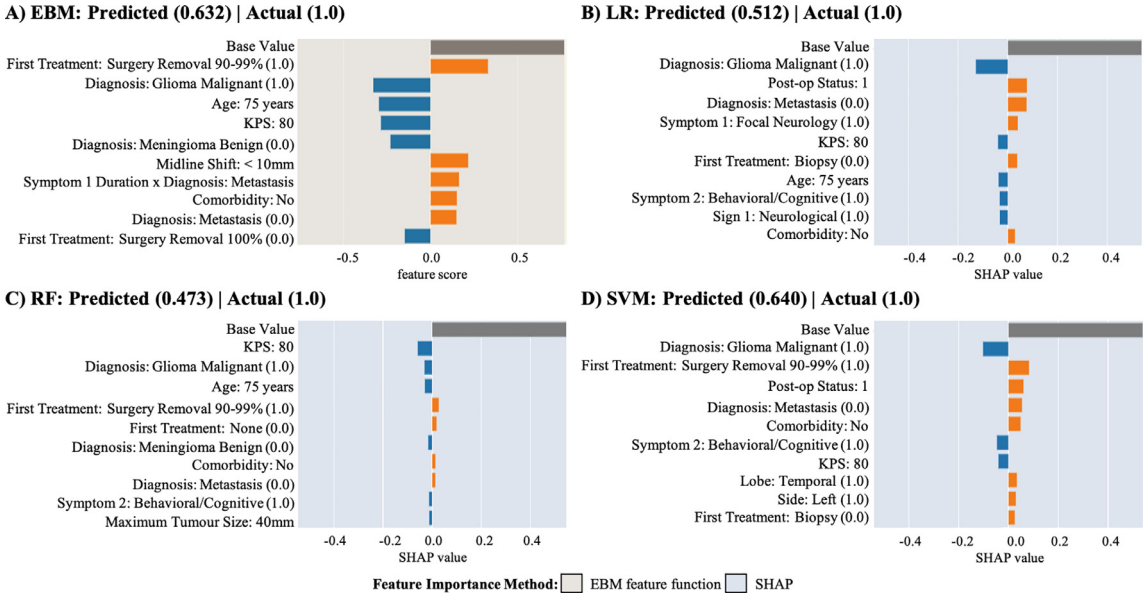

Fig. 4. Top 10 influential features for each model for a single test instance (patient). Negative (blue) features favour survival less than one year, and positive (orange) features favour survival greater than one year. The model baseline value (i.e. the average value of all predictions), is shown in grey. (A) Explainable Boosting Machine's (EBM) local feature importance based on individual feature functions. Both individual features and pairwise interactions (depicted as feature 1 x feature 2) are shown. Local feature importance for (B) Logistic Regression (LR) (C) Random Forest (RF) and (D) Support Vector Machine (SVM) determined by SHAP (SHapley Additive exPlanation). EBM, LR and SVM correctly classified the patient and RF incorrectly classified the patient.

As seen in Fig. 4, EBM also models feature interactions (depicted as *feature 1* x *feature 2*) thus further refining the interpretation of the model's prediction. For example, in the given test instance (Fig. 4A), the interaction between symptom 1 duration and metastatic brain tumour positively influences survival. Fig. 5 visualises this pairwise interaction as a heat map. We can see that having a metastatic tumour with a primary symptom less than ~8 weeks negatively influences survival greater than a year (score ~ -0.17), having a metastatic tumour with symptoms between 8 to 25 weeks has a minimal effect on survival (score ~ -0.01) and symptoms longer than 25 weeks have a positive effect on survival (score ~0.31). One interpretation is that more aggressive tumours may cause symptoms to present more rapidly, causing individuals to seek out medical attention faster and leading to a shorter recorded

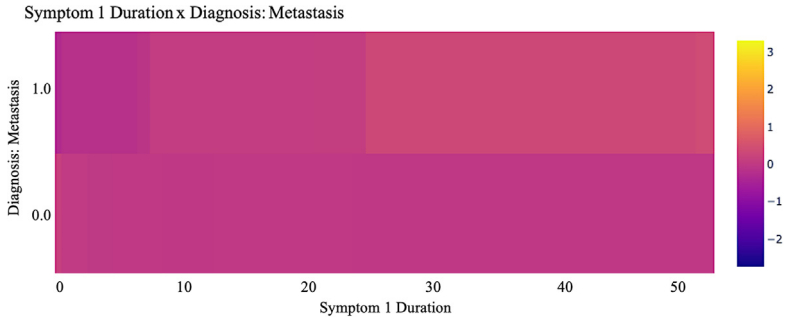

Fig. 5. Pairwise feature interaction, determined by the Explainable Boosting Machine model, between the features Symptom 1 Duration and Diagnosis: Metastasis. A score above zero positively influences survival greater than one year and a score below zero negatively influences survival greater than one year.

duration of the primary symptom. However, due to the aggressive nature of the tumour, this ultimately leads to a shorter survival time. This is in-line with EBM SHAP values (Fig. 3A), which show that shorter Symptom 1 durations negatively influence survival. Furthermore, according to the heatmap, *not* having a metastatic tumour positively influences survival despite the primary symptom duration. However, as we do not know the type of tumour the patient may have, a meaningful interpretation is restricted. Additional heat maps of pairwise interactions determined by EBM are shown in Appendix *A.4.2: Explainable Boosting Machine Pairwise Feature Interactions*.

4. Discussion

The results summarized in Table 1 show that RF outperformed all models, with a BAC of 78.9%. However, glass box models, EBM and LR, performed comparably well, and despite the lesser performance by BRL (72.6% BAC), the model achieved a specificity in a similar range of the other models (73%-75%). In comparison, EBM, LR, RF and SVM achieved a sensitivity above 80%, compared to BRL which attained a sensitivity of 70.6%. A high sensitivity ensures fewer false negatives, and thus correctly identifies individuals who will survive greater than one year, but a low specificity suggests the model may incorrectly identify individuals as surviving more than one year. For a patient, correctly predicting poor survival may be more important than correctly predicting long-term survival [47] and similar to the ML models, clinicians often overestimate survival time [48]. Interestingly, positive SHAP values, which favour survival greater than one year (see Fig. 3), reach higher values compared to negative SHAP values, thus having a greater impact on model predictions.

The innate interpretability of the glass box models, specifically EBM, may mitigate potential performance loss, especially in healthcare where model transparency is essential for its integration into and influence on clinical decision making [11]. The superior performance of EBMs has been reliably shown [16,49], and here we have demonstrated that it creates interpretable results without significant compromise in model performance. Additionally, EBM has the added benefit of considering feature interactions. Although LR innate interpretability can be assessed using the odds ratio, this metric is associated with the relative risk of an event, and does not meaningfully describe a feature's ability to classify subjects [50]. A recent systematic review found that prognostic models for predicting survival in glioblastoma patients using clinical, imaging and/or genomics data, achieved accuracies between 69-98% [51]. However, the authors highlight the importance of including secondary metrics, such as interpretability and ease of model use, that are relevant for the clinical deployment of models.

The findings of our feature importance analysis show that age, diagnosis, functional features (KPS and post-op performance status) and first treatment are ranked highly across all models. According to the experts, there is no gold standard of expected feature rankings, however the aforementioned variables would not be disputed and are reliably reported across clinical and epidemiological studies [43,44,45]. Furthermore, the experts suggest that multimorbidities and SIMD can effect survival, however these features are not as influential across models compared to other features. Nonetheless, by looking at SHAP global feature importance (Fig. 3), for all models the presence of a comorbidity (value of 1) negatively effects survival, while the absence of a comorbidity (value of 0) positively impacts survival. Furthermore, lower SIMD scores - corresponding to areas of greater deprivation - indicates poorer survival, while higher SIMD scores positively contribute to survival. Despite these features not being the most influential, we can see that the model's learn the expected feature contribution in-line with domain knowledge.

At both the global and local level, EBM and SHAP emphasized similar features. We can compare feature methods by looking at Fig. 2A, which illustrates EBM absolute global feature importance based on the learned feature functions, and Fig. 3A, which illustrates EBM feature importance based on SHAP values. Both methods highlight similar feature types with slightly different orderings, the most noticeable being *First Treatment: Surgery 100%*. EBM ranks this feature in position 6, while SHAP ranks this feature in position 15. However, given there are 58 different feature types, both methods produce similar results. Nonetheless, the difference in methodologies may be due to SHAP's assumption of feature independence. SHAP permutes feature values, sampled from the features marginal distribution, and makes predictions based on these permutations. When features are dependent, predictions may be based on unrealistic feature values leading to unreliable SHAP values. In comparison, EBM has the added benefit of considering feature interactions. Both methods are computationally expensive, making them somewhat slower than other interpretability methods. Although SHAP has the advantage of being model-agnostic, SHAP explanations can be manipulated to create intentionally misleading interpretations [52] and the credibility of post-hoc explanation methods continues to be debated [11,53,54]. A model which is interpretable by design, such as EBMs, may provide more faithful explanations, however this approach is model-specific making the comparison of interpretability between algorithms more difficult.

Finally, the various interpretability techniques highlight the technical differences in model construction. As illustrated in Fig. 2, RF and EBM both favour *Age* as the most informative feature and *Symptom 1 Duration* as moderately informative. In comparison, LR and SVM view *Age* as moderately informative, and do not consider *Symptom 1 Duration* to be important. Both RF and EBM are tree-based algorithms, and may favour the selection of features with many possible splits (e.g. continuous variables or categorical variables with high cardinalities) over variables with fewer splits [55].

LR and SVM find binary features to be of higher value, including *Comorbidity* and *Previous Medical History (PMH) of Cancer*. Interestingly, SVM finds a wider range of features to be informative, possible due to its effectiveness in high dimensional spaces, including the presence of a brain tumour on the left side. BRL also finds tumour side to be relevant (see first IF statement in Fig. 1) which may be overlooked by the other models. Previous studies have found that patients with a tumour on the right side had poorer quality of life than those with a tumour on the left side [56]. Furthermore, each ML model comes with its own unique advantages. For example, rule lists produce simple IF-THEN statements that are easy to interpret while tree-based algorithms such as RF and EBM, are beneficial for capturing interactions in the data. LR is quick to train and returns class probabilities while SVM is favourable for high-dimensional data. The advantages and disadvantages of various ML methods, along with the level of interpretability, are essential considerations when building high-stake prognostic models.

4.1. Limitations

To our knowledge, this is the first study to compare intrinsic and post-hoc interpretability methods for the assessment of predictive brain tumour survival models. Nonetheless, several important limitations remain. In the present study, only clinical prognostic factors were used for the prediction of survival. As previously mentioned, blood tests and molecular genetic alterations have been recognised as powerful prognostic and predictive markers in brain tumour survival [45,46,57]. The integration of the different data types may improve current survival predictions, but additional clinical testing may be costly, time consuming and increase patient burden (e.g., invasiveness), compared to readily available electronic patient records. In addition, the data used in this study was heterogeneous and an external validation dataset would be required to confirm the generalizability and reproducibility of our results. Furthermore, our dataset required imputation and discretization (for BRL). There is the potential for imputation to introduce bias into the data and the chosen discretisation method can influence the final results. However, multivariate imputation has been shown to outperform other techniques [58,59], while feature discretization was based on current literature and expert opinion, and is only relevant for BRL. Finally, rule-lists currently require categorical features and are limited to binary classification (although there has been some recent effort toward multi-class rule lists [60]). Extension of the rule list algorithm for multi-class classification or regression is an important next step for improving rule-list performance and constructing a competitive intrinsically interpretable rule list classifier.

5. Conclusion

In this study, we found that EBM performance was comparable to black box models, with RF outperforming EBM by less than 2% BAC (see Table 1), for the prediction of one year survival following a brain tumour diagnosis. EBMs provided valuable information on relevant prognostic factors (e.g. age, diagnosis, post-op performance status) and their interactions (e.g. Symptom 1 duration *x* diagnosis: metastasis). More generally, informative model features such as age, diagnosis, functional status (KPS and post-op performance status) and first treatment, were in-line with domain-knowledge and current literature. Our results also confirm that SHAP was beneficial for understanding model behaviour, although as explained in the literature, post-hoc explainability methods can be vulnerable to failures, so our results should also be interpreted with care. Interpretability is crucial for the implementation of ML algorithms in healthcare, and whether the model is innately interpretable or post-hoc methods are used, the validation and integration of such models into clinical practice is an important next step for improving patient outcomes in a trusted way.

Data availability

The data that support the findings of this study are available on request from the author P.B. (paul.brennan@ed.ac.uk) subject to relevant review board approval.

Declaration of Competing Interest

The authors declare no conflict of interest.

Funding

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

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.cmpb.2023.107482.

Appendix A

A.1. Data preprocessing

Initially, the raw brain tumour dataset contained 1283 patient records and 225 predictor variables. A preliminary exploratory analysis of the data found a large number of features ($n = 179$) had more than 60% of their entries not recorded. A large portion of these absent entries pertained to symptom and sign related features ($n = 110$). For example, the raw dataset included information on the first symptom type a patient presented with, up to the tenth symptom (e.g., Symptom 1, Symptom 2, Symptom 3, etc.). The same pattern occurred for signs. However, a patient may only present with one symptom thus leaving the remaining nine symptom features empty. Hence an entry for a feature may be absent, but this does not imply that the entry is truly missing. Furthermore, there was significant duplication amongst the symptom and sign features. For example, features included: Symptom 1, Symptom 1 - Headache (choice=Worse on coughing/bending), Symptom 1 - Headache (choice=Worse on waking), Symptom 1 - Headache (choice=Associated with nausea/vomiting), etc. This was repeated for symptoms 1 though 10. Taken together, this created the appearance of missing data, but in fact the empty entries are correctly missing. In the raw dataset, 91% of patient presented with at least one symptom, only 62% of patients presented with two symptoms, and less than 26% of patients presented with three symptoms. Given the amount of absent symptom and sign entries, only symptom 1 (i.e. the first symptom a patient presents with), symptom 2 and sign 1 were used as features in the final dataset, and the remaining features were removed ($n = 107$).

Furthermore, a large number of features pertained to treatment details, including features related to blood tests (e.g., baseline count of white blood cells, neutrophils, lymphocytes, platelets, etc.), chemotherapy protocol (e.g., date of chemotherapy start, chemotherapy drug, type of 2nd line chemotherapy, etc.) and other details (e.g., location of first imaging, clinician ordering CT [open access CT], contrast agent). However, if a patient did not receive a specific treatment (e.g., chemotherapy), a large number of features were again, correctly missing. The majority of treatment features, and other unrelated details (e.g. speciality referring for first scan, discharge destination), were sparse and removed from the dataset ($n = 98$).

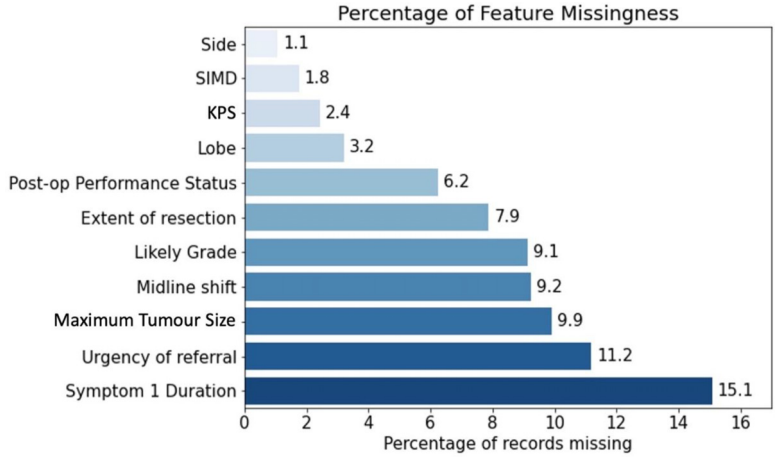

Fig. A1. Percentage of incomplete patient records in the final dataset. Only features with missing records are shown. 52% of subjects had no missing variables, 31% of subjects had one missing variable, 9% of subjects had two missing variables, 3% of subjects had three missing variables, 4% of subjects had four missing variables and <1% of subjects had five missing variables. See Appendix A.2: Description of Dataset Variables for description of features. SIMD: Scottish Index of Multiple Deprivation.

Additionally, a number of patients were removed due to incomplete records (i.e. a patient is missing more than 40% of the reduced predictor variables) ($n = 36$) or because of a lack of symptomatology information (i.e. a patient did not present with any symptoms or signs) ($n = 219$).

A.1.1. Missing data

Of the remaining 20 features, 11 had missing data as illustrated in Fig. A1. The feature Symptom 1 Duration, had the largest number of missing variables, with 15% of the records missing, followed by Urgency of Referral (11%) and Maximum Tumour Size (10%).

Through discussions with the experts, the 11 features were deemed missing at random. Although post-op performance status would only be relevant for patients who had surgery, and similarly, the extent of resection would only be relevant for those who had resective surgery (not biopsy), it was confirmed through dataset investigation, that these individuals did undergo surgery hence these variables were missing at random.

A.1.2. Feature correlations

As mentioned in Section 2.2, correlation between variables was determined through statistical testing and discussion with the experts. An appropriate measure of association was selected based on the type of variable we were assessing (i.e., continuous, ordinal or nominal). The six possible feature combinations and the chosen statistical test were chosen as follows:

- Continuous-continuous: Pearson correlation coefficient (r_{pcc})
- Continuous-ordinal: Kendall's tau (τ_b)
- Continuous-nominal: Point-biserial correlation coefficient (r_{pb})
- Ordinal-ordinal: Kendall's tau (τ_b)
- Ordinal-nominal: Point-biserial correlation coefficient (r_{pb})
- Nominal-nominal: Cramer's V (φ_c)

Note, Spearman correlation coefficient was also assessed as an alternative to Kendall's tau, but the results did not differ greatly.

Of the original 21 variables, we found *Morphology* to be highly correlated with *Tumour Type* and *Likely Grade* (φ_c = 0.77 and r_{pb} = -0.71, respectively). Due to the high correlation, *Morphology* was removed from the dataset and dataset preprocessing was re-run (see Section 2.2). *Tumour Type* and *Likely Grade* were combined into a single feature called *Diagnosis.* Tumour type refers to the kind of tumour a patient is diagnosed with, while likely grade is an indicator of how quickly a tumour is likely to grow or spread. For example, by definition, a glioblastoma is a grade IV (fast growing) glioma tumour. Despite *Tumour Type* and *Likely Grade* having a minimal correlation overall (φ_c = -0.38), to reduce collinearity in the data, the two features were concatenated. Tumour types were separated into benign (or low-grade) and malignant (or high-grade) categories (e.g. Glioma Benign and Glioma Malignant).

Additionally, we found *Post-op Performance Status* to be highly correlated with *Extent of Resection* (EOR) and *First Treatment* (τ_b = -0.64 and r_{pb} = 0.62, respectively). EOR refers to the amount of cancerous cells removed during surgery (e.g. 90-99%) and is only relevant if the first treatment a patient receives is surgery, which is not always the case. For example, if a person received chemotherapy as their first treatment, an entry for *Extent of Resection* would be absent, but the data was correctly missing. Thus we integrated EOR information into First Treatment (e.g. Surgery 100%, Surgery 90-99%) to create a more informative feature - the feature name remained *First Treatment*. Following feature concatenation, *Post-op Performance Status* and the new variable *First Treatment* had a minimal correlation (r_{pb} = -0.29). Hence the final dataset contained 1028 patient records and 18 predictor variables and one dependent variable, namely patient survival in days. Of the final 18 features, all features had correlations below 0.5, with the exception of *Diagnosis* and *Post-op Performance Status*, which had a correlation of 0.56.

A.1.3. Feature discretisation

For the BRL model, continuous variables were discretised into meaningful categories based on current literature and expert opinion. As the overall aim of these models is to be maximally interpretable, manual discretisation was based on well-defined cut points was used. The variables were discretized as follows:

- Age: < 45, 45-59, 60-74 and 75+
  - Symptom 1 Duration: <2 weeks, 3-4 weeks, 5-8 weeks, 9-20 weeks, 21+ weeks
  - Maximum Tumour Size: <20 mm, 21-40 mm, 41-60 mm, 60+ mm

Fig. A2 provides additional information on the distribution of the continuous features. SeeTable A2 for continuous feature summary statistics.

A.1.4. Other feature preprocessing steps

Finally, we briefly discuss some of the salient features below and some final processing of the features.

KPS: This is a standard way of assessing a patient's ability to perform everyday tasks [32]. The scale is a 'gold standard' in clinical oncology and is commonly used to determine a cancer patient's

Table A1
Overview of dataset variables including their descriptions, value and percentage of each value present in the final dataset. Imputation was performed over the whole dataset thus providing an approximation of the values present in the individual train test splits.

Name	Description	Value	Proportion (%)
Age	the age of a patient	16-97	
Sex	the sex of the patient	Male	50.3
		Female	49.7
History of Cancer	whether the patient has a past medical history of cancer	Yes	18.1
		No	81.9
Comorbidity	the presence of another illness or disease occurring in a patient	Yes	47.5
		No	52.5
Scottish Index of Multiple Deprivation (SIMD)	a measure of deprivation of the area a patient lives from most deprived (ranked 1) to least deprived (ranked 5)	1	13.8
		2	22.9
		3	21.6
		4	18.5
		5	23.2
Karnofsky Performance Score (KPS)	a common measure in oncology to assess the functional state of a patient	100	36.8
		90	29.1
		80	15.6
		$\leq$ 70	18.5
Symptom 1	the first symptom type a patient presented with (reported by the patient)	Focal Neurology	34.6
		Headache	28.4
		Behavioural/Cognitive	16.8
		Fits/Faints/Falls	16.8
		Other/Non-specific	2.4
		Non-specific Neurological	1.0
Symptom 1 Duration	the length of time of a patient's first symptom	0 – 52 weeks	
Symptom 2	the second symptom type a patient presented with (reported by the patient)	Focal Neurology	31.4
		No Symptoms	30.1
		Behavioural/Cognitive	19.1
		Fits/Faints/Falls	8.9
		Headache	6.7
		Other/Non-specific	3.9
Sign 1	the first sign type a patient presented with (observed by the physician)	No Signs	42.1
		Neurological	36.4
		Cognitive	15.2
		Cranial Nerve	5.1
		Other	0.7
		Behavioural	0.5
Urgency of Referral	the patient's urgency of referral from primary care	Emergency	61.5
		Suspicion of Cancer (within 2 weeks)	22.8
		Soon (up to 3-4 weeks)	5.9
		Routine (up to 12 weeks)	9.8
Diagnosis (or Tumour Type)	the type of brain tumour a patient was diagnosed with	Glioma Malignant	45.6
		Metastasis	18.9
		Meningioma Benign	11.9
		Glioma Benign	7.7
		Rare Tumour Benign	5.5
		Lymphoma Benign	4.4
		Meningioma Malignant	4.2
		Hemangioblastoma Benign	1.2
		Rare Tumour Malignant	$<$0.01
Max Size	a measure of the tumour size	1-120	
Side	the side of the brain the tumour is located	Left	42.2
		Right	40.9
		Both Left and Right	11.4
		Midline	5.5
Lobe	the lobe where the tumour is located	Frontal	36.8
		Temporal	21.4
		Parietal	14.2
		Multiple	12.1
		Cerebellar	7.0
		Brainstem	4.3
		Occipital	4.2
Midline Shift	a measure of the tumour's horizontal shift from the mid (centre) line	0	40.9
		$<$ 5mm	28.2
		5-10mm	19.3
		$>$ 10mm	11.6
First Treatment	the type of first cancer treatment	Surgery Removal 100%	15.2
		Surgery Removal 90-99%	23.2
		Surgery Removal 50-89%	8.4
		Surgery Removal $<$ 50%	5.1
		Biopsy	16.6
		Radiotherapy	5.4
		Chemotherapy	0.9
		Other (e.g. steroids)	2.6
		No Treatment	22.6

(*continued on next page*)

Table A1 (*continued*)

Name	Description	Value	Proportion (%)
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	0	29.6
		1	27.0
		2	7.6
		3	2.7
		4	1.5
		5	<0.1
		No Surgery	31.5
1-Year Survival		> 1-Year	43
		≤ 1-Year	57

Table A2
Description of continuous variables including mean, and in parenthesis standard deviation (SD), median and mode.

	Mean (SD)	Median	Mode
Age	59 (15)	61	54
Maximum Tumour Size	39 (18)	39	40
Symptom 1 Duration	11 (17)	4	0

expected tolerance to treatments (e.g. chemotherapy). The scores ranges from 0 (dead) to 100 (normal) and is scored in deciles, although the values are ordinal (see Appendix A.2: Table A3 for the original definition of the KPS). This means that a value assigned to a patient is based on a ranking but the numerical value associated with this rank is not meaningful. Thus the difference between the values 70 and 90 is not equivalent to the difference between the values 40 and 60. Furthermore, the KP scale may be subject to bias [61]. A patient's KPS is determined by clinicians, and when compiling a dataset this can result in inter-observer subjectivity [62]. To reduce the bias associated with the KPS and based on the advice of the consulting clinical experts, values of 70 and below were aggregated due to their negative association with survival [63] (a KPS of 70 reflects someone who can 'care for self, but who is unable to carry on normal activity or to do active work'). KPS of 80 and above remained separate allowing for a more fine-grained analysis of the values associated with survival.

Symptom 1. : A symptom is observed by the patient themselves (subjective) and is often what drives a patient to consult a physician. Symptom 1 refers to the first symptom a patient presents with. The symptom data in the raw dataset had a high cardinality of 37 different symptom types, with many of these types pertaining to a small number of patients. Thus we decided to group symptom types into six overarching categories – e.g. Headache, Fits/Faints/Falls and Behavioural/Cognitive – based on work by Ozama et al. [64] to create a more homogeneous set of symptom types. An outline of the symptom groupings are summarised in Appendix A.2: Table A4.

Sign 1. : A sign is observed by a physician (objective). Sign 1 refers to the first sign a patient presents with. The sign data in the raw dataset also had a high cardinality (26 different types), thus the data was additionally grouped into six larger domains – e.g. neurological and cognitive– based on the advice of the consulting clinical experts (see Appendix A.2: Table A5). Although all patients in the final reduced dataset presented with at least one symptom, 43% of patients did not present with any signs.

Diagnosis (Tumour Type): Brain tumours are broadly named based on the type of normal cell that they most resemble, and their location in the brain [65]. In the raw dataset the tumour types had a high cardinality with many entries referring to the same general tumour type (e.g. meningioma suprasellar and meningioma at cerebellopontine [CP] angle). Tumour types that appeared in less than 10 patients were grouped into a "Rare Tumour" category. Additionally, the tumour type may be benign (i.e. grade I/II), or malignant (i.e. grade III/IV). In the final dataset, the brain tumour types were reorganised based on type and malignancy (e.g. Glioma Benign, Glioma Malignant), and reduced to a cardinality of 9.

First Treatment: The first type of cancer treatment a patient receives is based on the presumed type based on imaging, location of the tumour, and the patient's overall health (e.g. KPS ≤ 70). Surgery, for example, may be the only treatment necessary depending on the grade of the tumour and extent of resection. Information on extent of resection was included in the first treatment types (e.g. Surgery 100%, Surgery 90-99%, Biopsy, etc.). Other treatment types include radiotherapy and chemotherapy.

A.2. Description of Dataset Variables

A detailed overview of each feature, including their descriptions, values and proportion, is provided in Table A 1.

Summary statistics of the continuous features, including mean, median and mode, is provided in Table A2.

A description of the Karnofsky performance status is provided in Table A3.

A summary of symptom domain groupings is provided in Table A4.

A summary of sign domain groupings is provided in Table A5.

The Eastern Cooperative Oncology Group Performance Status Scale is provided in Table A6.

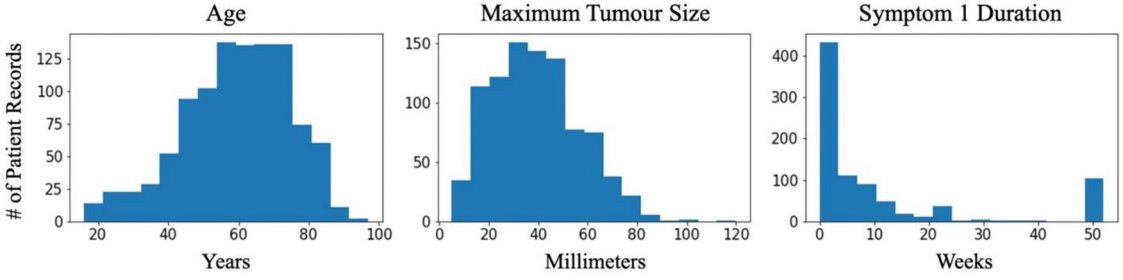

Fig. A2. Frequency histogram based on 15 bins for all continuous variables in the dataset.

Table A3
The original description of the Karnofsky performance status given by Karnofsky and Burchenal [32].

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 and care for most personal needs; varying amount of assistance 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 personal 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; hospital admission is indicated although death not imminent
	20	Very sick; hospital admission necessary; active supportive treatment necessary.
	10	Moribund; fatal processes progressing rapidly.
	0	Dead.

Table A4
Symptom domain classifications based on Ozawa et al. [64], with examples of symptom types in the brain tumour dataset.

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 A5
Sign domain classifications based on clinical expertise of some of the current authors. All examples are from the Brain Tumour dataset.

Group	Sign Domain	Sign Examples
1	No Signs	No Signs
2	Behavioral	Behaviour signs anxiety (e.g. fast speech, tremor, voices anxiety, crying)
		Behaviour signs depression (e.g. voices low mood, crying)
		Behaviour (withdrawn/apathetic) - not depressed
		Behaviour (aggressive/paranoid) - not anxious
3	Cognitive	Cognitive - problems performing tasks (e.g. calculation, planning, VF)
		Cognitive - problems with memory (forgetfulness)
		Cognitive - reduced conscious level/drowsiness (reduced GCS)
		Cognitive - other non-specific confusion
4	Neurological	Ataxia, vertigo, vision problems
		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 A6
Description of a patient's performance status (or functional state) developed by the Eastern Cooperative Oncology Group [66].

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 A7
The range of values the hyperparameters could take during hyperparamter optimisation within the nested cross validation.

Model	Hyperparameter	Search Distribution
Bayesian Rule List	listlengthprior	5, 10
	maxcardinality	2, 3
Explainable Boosting Machine	min_samples_leaf	1, 2, 4, 8
	max_leaves	2, 3, 5
	learning_rate	0.01, 0.05
Logistic Regression	C (regularization parameter)	2^{-4}, 2^{-2},..., 2^{4}
Random Forest	max_depth	10, 20, 30, None
	min_samples_leaf	1, 2, 4, 8
	min_samples_split	2, 5, 10
Support Vector Machine	C (regularization parameter)	2^{-4}, 2^{-2},..., 2^{4}
	gamma	scale, auto

A.3. Hyperparameter Tuning

A description of the hyperparameter space is provided in Table A7.

A.4. Modelling Results

A.4.1. Local Feature Importance

Fig. A3 illustrates the 10 most influential features for the classification of a single patient from the EBM, LR, RF and SVM models.

A.4.2. Explainable Boosting Machine Pairwise Feature Interactions

Pairwise feature interactions, determined by EBM, are visualized as heatmaps in Figure A4.

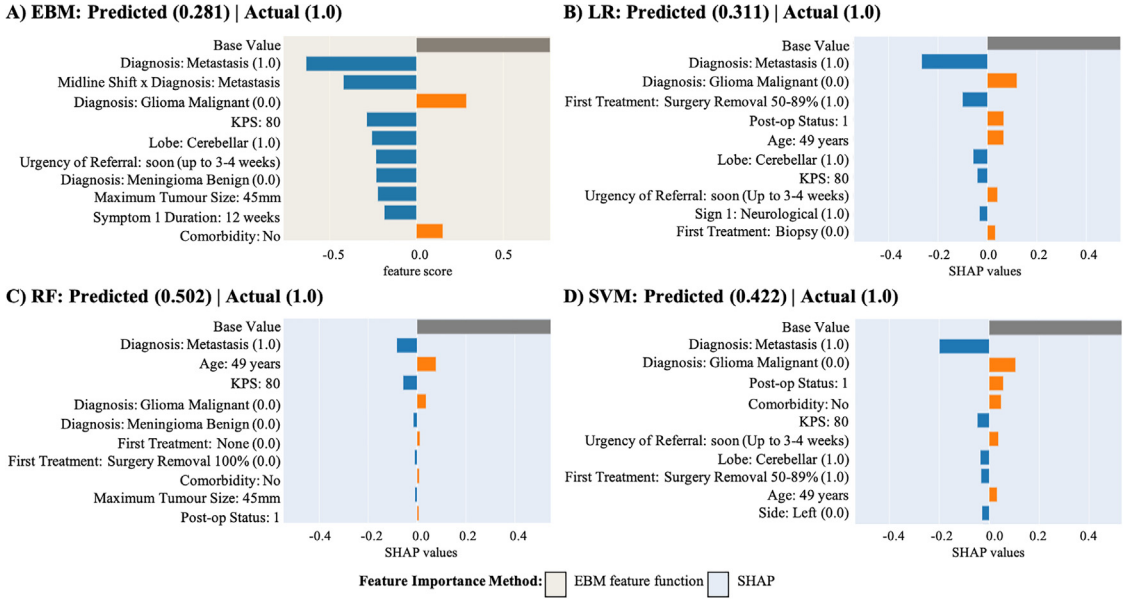

Fig. A3. Top 10 influential features for each model for a single test instance (patient). Negative (blue) features favour survival less than one year, and positive (orange) features favour survival greater than one year. The model baseline value (i.e. the average value of all predictions), is shown in grey. (A) Explainable Boosting Machine (EBM) local feature importance based on individual feature functions. Both individual features and pairwise interactions (depicted as feature 1 x feature 2) are shown. Local feature importance for (B) Logistic Regression (LR) (C) Random Forest(RF) and (D) Support Vector Machine (SVM) determined by SHAP (SHapley Additive exPlanations). RF correctly classified the patient and EBM, LR and SVM incorrectly classified the patient.

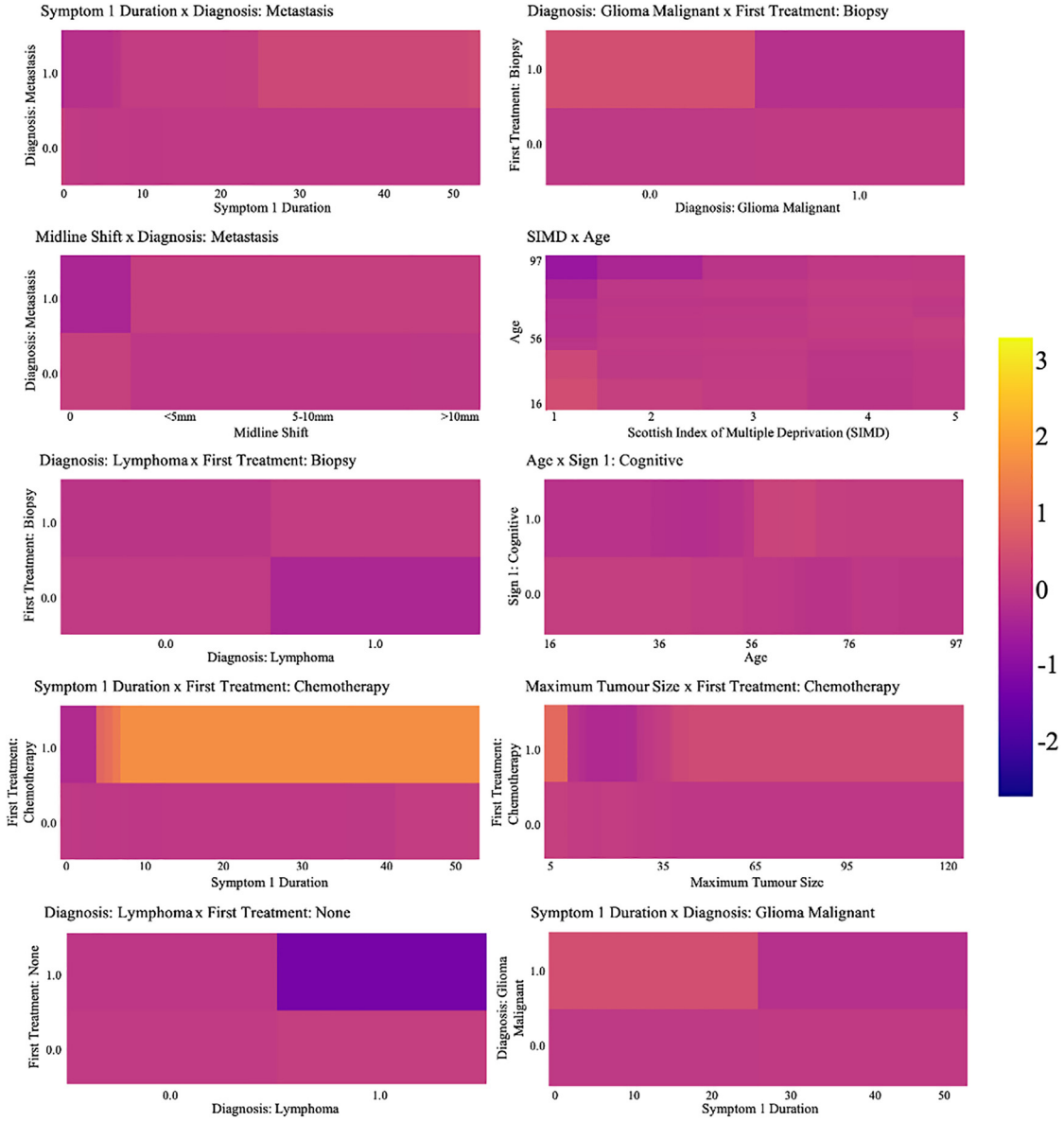

Fig. A4. Heat maps of all pairwise feature interactions determined by explainable boosting machine. A value above zero favors survival greater than one year. Abbreviations: SIMD: Scottish Index of Multiple Deprivation.

Appendix B

Additional BRL point estimates provided to the experts for review are shown in Figs. B1,B2,B3 and B4.

	Condition		Probability	Credible Interval	Support
IF	First Treatment: None **AND** Diagnosis: Glioma Malignant	THEN chance of survival > 1 Year	2.8%	(0.3%-7.6%)	75
ELSE IF	Diagnosis: Metastasis	THEN chance of survival > 1 Year	33.8%	(26.6%-41.2%)	158
ELSE IF	Midline Shift: 0 **AND** KPS 100	THEN chance of survival > 1 Year	95.1%	(91.0%-98.0%)	136
ELSE IF	Symptom 1: Behavioral/Cognitive **AND** Diagnosis: Glioma Malignant	THEN chance of survival > 1 Year	31.1%	(20.3%-43.2%)	94
ELSE IF	First Treatment: Surgery Removal 100%	THEN chance of survival > 1 Year	97.0%	(91.8%-99.6%)	56
ELSE IF	Maximum Tumour Size < 20mm	THEN chance of survival > 1 Year	95.7%	(84.6%-99.9%)	22
ELSE IF	Age < 45	THEN chance of survival > 1 Year	92.3%	(83.8%-97.8%)	49
ELSE IF	First Treatment: Biopsy **AND** Diagnosis: Glioma Malignant	THEN chance of survival > 1 Year	26.2%	(16.1%-37.9%)	60
ELSE IF	Post-op Status: 1	THEN chance of survival > 1 Year	76.1%	(65.6%-85.2%)	58
ELSE IF	Side: Right	THEN chance of survival > 1 Year	38.7%	(27.1%-51.0%)	58
ELSE IF	Post-op Status : 2	THEN chance of survival > 1 Year	79.4%	(64.5%-91.0%)	5
ELSE		THEN chance of survival > 1 Year	32.5%	(19.1%-47.6%)	51

Fig. B1. BRL-point estimate with a balanced accuracy of 76.5% obtained from the first random seed.

	Condition		Probability	Confidence Interval	Support
IF	Diagnosis: Glioma Benign	THEN chance of survival > 1 Year	97.0%	(91.8%-99.6%)	65
ELSE IF	Age < 45	THEN chance of survival > 1 Year	91.5%	(85.5%-96.0%)	101
ELSE IF	Diagnosis: Metastasis	THEN chance of survival > 1 Year	29.8%	(22.8%-37.3%)	149
ELSE IF	First Treatment: Surgery Removal 100%	THEN chance of survival > 1 Year	93.3%	(86.7%-97.8%)	74
ELSE IF	KPS <=70 **AND** Sex: Female	THEN chance of survival > 1 Year	9.3%	(3.1%-18.2%)	54
ELSE IF	Post-op Status: 2 **AND** Diagnosis: Glioma Malignant	THEN chance of survival > 1 Year	40.4%	(31.2%-49.9%)	27
ELSE IF	Diagnosis: Meningioma Benign	THEN chance of survival > 1 Year	95.7%	(88.5%-99.5%)	44
ELSE IF	Midline Shift: 0 **AND** KPS: 100	THEN chance of survival > 1 Year	85.0%	(72.6%-94.1%)	47
ELSE IF	Age 45-59 **AND** Post-op Status: 1	THEN chance of survival > 1 Year	69.4%	(53.7%-83.1%)	41
ELSE IF	First Treatment : Surgery Removal 90-99%	THEN chance of survival > 1 Year	68.6%	(52.5%-82.6%)	54
ELSE IF	Diagnosis : Glioma Malignant	THEN chance of survival > 1 Year	6.1%	(2.3%-11.5%)	141
ELSE		THEN chance of survival > 1 Year	68.8%	(52.0%-83.3%)	25

Fig. B2. BRL-point estimate with a balanced accuracy of 76.2% obtained from the first random seed.

	Condition		Probability	Confidence Interval	Support
IF	Age < 45	THEN chance of survival > 1 Year	92.5%	(87.7%-96.2%)	144
ELSE IF	First Treatment: Surgery Removal 90-99% **AND** KPS: 100 **AND** Urgency of referral: Emergency	THEN chance of survival > 1 Year	83.8%	(70.5%-93.6%)	35
ELSE IF	Age: 60-74 **AND** Comorbidity: Yes **AND** Diagnosis: Glioma Malignant	THEN chance of survival > 1 Year	17..3%	(9.9%-26.2%)	80
ELSE IF	KPS: <=70 **AND** Diagnosis: Glioma Malignant	THEN chance of survival > 1 Year	9.1%	(3.5%-17.0%)	62
ELSE IF	Symptom 1 Duration: 21+ weeks **AND** Post-op Performance Status: 1	THEN chance of survival > 1 Year	9.3%	(3.1%-18.2%)	25
ELSE IF	First Treatment: Biopsy	THEN chance of survival > 1 Year	50.7%	(39.2%-62.2%)	74
ELSE IF	Diagnosis: Metastasis	THEN chance of survival > 1 Year	23.8%	(16.8%-31.6%)	130
ELSE IF	First Treatment: None **AND** Diagnosis : Glioma Malignant	THEN chance of survival > 1 Year	12.9%	(3.8%-26.5%)	29
ELSE IF	First Treatment: Surgery Removal 100%	THEN chance of survival > 1 Year	97.1%	(92.1%-99.6%)	61
ELSE IF	Symptom 1: Fits/Faints/Falls	THEN chance of survival > 1 Year	70.6%	(54.5%-84.4%)	30
ELSE IF	Urgency of referral: routine (up to 12 weeks)	THEN chance of survival > 1 Year	96.2%	(86.3%-99.9%)	25
ELSE IF	KPS: <=70	THEN chance of survival > 1 Year	96.2%	(2.1%-38.5%)	11
ELSE IF	Sign 1 : No Signs	THEN chance of survival > 1 Year	72.3%	(58.9%-84.0%)	44
ELSE		THEN chance of survival > 1 Year	62.3%	(50.7%-73.3%)	72

Fig. B3. BRL-point estimate with a balanced accuracy of 73.3% obtained from the first random seed.

	Condition		Probability	Confidence Interval	Support
IF	KPS: <=70 **AND** Symptom 2: Focal Neurology	THEN chance of survival > 1 Year	13.2%	(5.6%-23.4%)	49
ELSE IF	Diagnosis: Metastasis	THEN chance of survival > 1 Year	33.3%	(25.8%-41.3%)	140
ELSE IF	First Treatment: Surgery Removal 100%	THEN chance of survival > 1 Year	95.0%	(90.0%-98.3%)	98
ELSE IF	Symptom 1: Behavioral/Cognitive **AND** KPS: <=70	THEN chance of survival > 1 Year	12.2%	(4.2%-23.7%)	39
ELSE IF	Age: < 45	THEN chance of survival > 1 Year	93.5%	(88.1%-97.3%)	104
ELSE IF	First Treatment: None **AND** Diagnosis: Glioma Malignant	THEN chance of survival > 1 Year	2.1%	(0.1%-7.5%)	47
ELSE IF	Diagnosis: Glioma Benign	THEN chance of survival > 1 Year	86.2%	(71.8%-96.0%)	24
ELSE IF	Midline shift: 0 **AND** KPS: 100	THEN chance of survival > 1 Year	94.9%	(88.1%-98.9%)	53
ELSE IF	First Treatment: Biopsy **AND** Comorbidity: Yes	THEN chance of survival > 1 Year	20.8%	(10.7%-33.3%)	47
ELSE IF	First Treatment: None	THEN chance of survival > 1 Year	52.2%	(32.2%-71.8%)	23
ELSE IF	Diagnosis: Glioma Malignant	THEN chance of survival > 1 Year	50.0%	(41.9%-58.1%)	147
ELSE		THEN chance of survival > 1 Year	94.2%	(86.5%-98.8%)	51

Fig. B4. BRL-point estimate with a balanced accuracy of 61.2% obtained from the first random seed.

References

[1] M.T.C. Poon, C.L.M. Sudlow, J.D. Figueroa, P.M. Brennan, Longer-term (≤ 2 years) survival in patients with glioblastoma in population-based studies pre-and post-2005: a systematic review and meta-analysis, Sci. Rep. 10 (2020) 1–10.

[2] E.B. Claus, et al., Survival and low-grade glioma: the emergence of genetic information, Neurosurg. Focus 38 (2015) E6.

[3] R. Jain, et al., Outcome prediction in patients with glioblastoma by using imaging, clinical, and genomic biomarkers: focus on the nonenhancing component of the tumor, Radiology 272 (2014) 484–493.

[4] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17.

[5] P. Fulop, A. Manataki, A. Agachi, P. Pop, Predicting survival after surgery for brain tumour patients: A machine learning study on clinical data and molecular data, in: *Proceedings of the AI for Social Good workshop, 7th International Conference on Learning Representations*, 2019.

[6] M.T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144.

[7] J.T. Senders, et al., An online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning, Neurosurgery 86 (2020) E184–E192.

[8] L.J. Wei, The accelerated failure time model: a useful alternative to the cox regression model in survival analysis, Stat. Med. 11 (1992) 1871–1879.

[9] P.I. D'Urso, Letter: an online calculator for the prediction of survival in glioblastoma patients using classical statistics and machine learning, Neurosurg 87 (2020) E273–E274.

[10] M.A. Ahmad, C. Eckert, A. Teredesai, G. McKelvey, Interpretable machine learning in healthcare, IEEE Intell. Info. Bull. 19 (2018) 1–7.

[11] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nat. Mach. Intell. 1 (2019) 206–221.
[12] T. Miller, Explanation in artificial intelligence: insights from the social sciences, Artif. Intell. 267 (2019) 1–38.
[13] D.V. Carvalho, E.M. Pereira, J.S. Cardoso, Machine learning interpretability: a survey on methods and metrics, Electronics 8 (2019) 832.
[14] Molnar, C. *Interpretable Machine Learning*. (2019). Available online: https://christophm.github.io/interpretable-ml-book/(accessed on 10 August 2021).
[15] B. Letham, C. Rudin, T.H. McCormick, D. Madigan, Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model, Ann. Appl. Stat. 9 (2015) 1350–1371.
[16] Nori, H., Jenkins, S., Koch,P. & Caruana, R. InterpretML: A unified framework for machine learning interpretability. Preprint at https://arxiv.org/abs/1909.09223 (2019).
[17] S. Menard, *Applied logistic regression analysis* No. 106, Sage, 2002.
[18] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32.
[19] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (1999) 293–300.
[20] S.M. Lundberg, S. Lee, A unified approach to interpreting model predictions, Adv. Neural Inf, Process Syst. 30 (2017).
[21] G.S. Collins, J.B. Reitsma, D.G. Altman, K.G. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement, BR. J. Surg. 102 (2015) 148–158.
[22] Rossum, G. V. & Drake, L. F. Python language reference manual. (2003).
[23] W. McKinney, et al., Data structures for statistical computing in python, in: Proceedings of the 9th Python in Science Conference, 445, 2010, pp. 51–56.
[24] T.E. Oliphant, A guide to NumPy, Trelgol Publishing, 2006.
[25] J.D. Hunter, Matplotlib: A 2d graphics environment, Comput. Sci. Eng. 9 (2007) 90–95.
[26] F. Pedregosa, et al., Scikit-learn: machine learning in Python, JMLR 12 (2011) 2825–2830.
[27] Alvarez-Melis, D. & Jaakkola, T. S. On the robustness of interpretability methods. Preprint at https://arxiv.org/abs/1806.08049 (2018).
[28] L. Antwarg, R.M. Miller, B. Shapira, L. Rokach, Explaining anomalies detected by autoencoders using Shapley additive explanations, Expert Syst. Appl. 186 (2021) 115736.
[29] M.E. Tipping, Sparse Bayesian learning and the relevance vector machine, JMLR 1 (2001) 211–244.
[30] S.V. Buuren, K. Groothuis-Oudshoorn, MICE: Multivariate imputation by chained equations in R, J. Stat. Softw. 45 (2011) 1–67.
[31] J.W. Grzymala-Busse, J Stefanowski, Three discretization methods for rule induction, Int. J. Intell. Syst. 16 (2001) 29–38.
[32] D.A. Karnofsky, W.H. Abelmann, L.F. Craver, J.H. Burchenal, The use of the nitrogen mustards in the palliative treatment of carcinoma. with particular reference to bronchogenic carcinoma, Cancer 1 (1948) 634–656.
[33] S. Varma, R Simon, Bias in error estimation when using cross-validation for model selection, BMC Bioinform 7 (2006) 91.
[34] M. Hossin, M.N. Sulaiman, A review on evaluation metrics for data classification evaluations, IJDKP 5 (2015) 1.
[35] A.P. Bradley, The use of the area under the ROC curve in the evaluation of machine learning algorithms, Pattern Recognit. 30 (1997) 1145–1159.
[36] C. Borgelt, An implementation of the FP-growth algorithm, in: Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations, 2005, pp. 1–5.
[37] K.W. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97–109.
[38] L.E. Eberly, G. Casella, Estimating Bayesian credible intervals, J. Stat. Plan. Inference 112 (2003) 115–132.
[39] S.N. Wood, Generalized additive models: an introduction with R, CRC press, 2017.
[40] J. Park, I.W. Sandberg, Universal approximation using radial-basis-function networks, Neural Comput. 3 (1991) 246–257.
[41] A.E. Roth, The Shapley value: essays in honor of Lloyd S. Shapley, Cambridge University Press, 1988.
[42] M. Szumilas, Explaining odds ratios, J. Can. Acad. Child Adolesc. 19 (2010) 227.
[43] S.R. Dehcordi, Survival prognostic factors in patients with glioblastoma: our experience, J. Neurosurg. Sci. 56 (2012) 239–245.
[44] H. Gittleman, Survivorship in adults with malignant brain and other central nervous system tumor from 2000–2014, Neuro-Oncol 20 (2018) vii6–vii16.
[45] S. Lapointe, A. Perry, N.A. Butowski, Primary brain tumours in adults, The Lancet 392 (2018) 432–446.
[46] S. Podnar, et al., Diagnosing brain tumours by routine blood tests using machine learning, Sci. Rep. 9 (2019) 1–7.
[47] P.A. Glare, C.T. Sinclair, Palliative medicine review: prognostication, J. Palliat. Med. 11 (2008) 84–103.
[48] S. Cheon, et al., The accuracy of clinicians' predictions of survival in advanced cancer: a review, Ann. Palliat. Med. 5 (2016) 22–29.
[49] R. Caruana, et al., 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, 2015, pp. 1721–1730.
[50] M.S. Pepe, H. Janes, G. Longton, W. Leisenring, P. Newcomb, Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker, Am. J. Epidemiol. 159 (2004) 882–890.
[51] I.A. Tewarie, Survival prediction of glioblastoma patients—Are we there yet? A systematic review of prognostic modeling for glioblastoma and its clinical potential, Neurosurg. Rev. 44 (2021) 2047–2057.
[52] D. Slack, S. Hilgard, E. Jia, S. Singh, H. Lakkaraju, Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods, in: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 180–186.
[53] T. Laugel, M. Lesot, C. Marsala, X. Renard, M. Detyniecki, The dangers of post-hoc interpretability: Unjustified counterfactual explanations, in: Proceedings of the 28th International Joint Conference on Artificial Intelligence, AAAI, 2019, pp. 2801–2807.
[54] B. Dimanov, U. Bhatt, M. Jamnik, A. Weller, You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods, SafeAI@ AAAI (2020) 63–73.
[55] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and regression trees, Routledge, 2017.
[56] J. Salo, A. Niemelä, M. Joukamaa, J. Koivukangas, Effect of brain tumour laterality on patients' perceived quality of life, JNNP 72 (2002) 373–377.
[57] E. Gray, et al., Health economic evaluation of a serum-based blood test for brain tumour diagnosis: exploration of two clinical scenarios, BMJ Open 8 (2018) e017593.
[58] G. Ambler, R.Z. Omar, P Royston, A comparison of imputation techniques for handling missing predictor values in a risk model with a binary outcome, Stat. Methods Med. Res. 16 (2007) 277–298.
[59] C.T. Tran, M. Zhang, P. Andreae, B. Xue, L.T Bui, An effective and efficient approach to classification with incomplete data, Knowl. Based Syst. 154 (2018) 1–16.
[60] H.M. Proenca, M. van Leeuwen, Interpretable multiclass classification by MDL based rule lists, Inf. Sci. 512 (2020) 1372–1393.
[61] D. Frappaz, et al., Assessment of Karnofsky (KPS) and WHO (WHO-PS) performance scores in brain tumour patients: The role of clinician bias, Support. Care Cancer 29 (2020) 1–9.
[62] J.B. Sørensen, M. Klee, T. Palshof, H.H. Hansen, Performance status assessment in cancer patients. an inter-observer variability study, Br. J. Cancer 67 (1993) 773–775.
[63] K. Chaichana, S. Parker, A. Olivi, A. Quiˇnones-Hinojosa, A proposed classification system that projects outcomes based on preoperative variables for adult patients with glioblastoma multiforme, J. Neurosurg. 112 (2010) 997–1004.
[64] M. Ozawa, The usefulness of symptoms alone or combined for general practitioners in considering the diagnosis of a brain tumour: a case-control study using the clinical practice research database (CPRD) (2000-2014), BMJ Open 9 (2019) e029686.
[65] The Brain Tumour Charity. Adult brain tumour types. 2020. Available online: https://www.thebraintumourcharity.org/brain-tumour-diagnosis-treatment/types-of-braintumour-adult (accessed on 1 August 2021).
[66] M.M. Oken, et al., Toxicity and response criteria of the eastern cooperative oncology group, Am. J. Clin. Oncol. 5 (1982) 649–656.