PREDICTING SURVIVAL AFTER SURGERY FOR BRAIN TUMOUR PATIENTS: A MACHINE LEARNING STUDY ON CLINICAL DATA AND MOLECULAR DATA

Patric Fulop & Areti Manataki
The University of Edinburgh
{patric.fulop,a.manataki}@ed.ac.uk

Alex Agachi
Empiric Capital
alex.agachi@gmail.com

Paul Pop
Neurolaboratories Limited
paul@neurolabs.eu

1 INTRODUCTION

There has been significant progress in applying machine learning and deep learning models in critical domains such as brain and skin cancer prevention and detection (Haenssle et al., 2018; Bakas et al., 2018), but also in prediction of overall survival. In 2019, brain tumours remain one of the most intractable types of cancers. As difficult to treat surgically as with radiotherapy or chemotherapy, survival expectancy is bleak. For Glioblastoma Multiforme (GBM), expected survival time from diagnosis is 3 months without treatment and 14 months with the best treatments available (Louis et al., 2016; Gallego, 2015). Most of the successful machine learning modelling revolves around using imaging data and state-of-the-art deep learning models for segmentation and survival prediction. For example the BRATS challenge (Bakas et al., 2018; Akkus et al., 2017) uses mpMRI and MRI data and the recent work of Lao et al. (2017) predicts overall survival of GBM by also applying transfer learning methods on radiomics data. Other studies also target genetic data for classifying brain tumours (Panca & Rustam, 2017).

However, there have been very few studies done on applying supervised and unsupervised methods on clinical data alone, in order to predict brain cancer survival. Although complete genomic data could prove richer, it is rarely available in clinical practice and neuro-oncologists still rely mostly on clinical and molecular data to make decisions. A crucial question for brain cancer patients and their families upon diagnosis is, unfortunately, how long they have left to live. In partnership with a neuro-oncological laboratory in Europe, we set out to apply several machine learning methods to one of the largest and most diverse, brain tumor databases in the world. The main goals for our study were two-fold:

- Apply clustering methods to identify patterns in our clinical and molecular data in order to provide useful insights to the clinicians.
- Predict patient survival after surgery using black-box classifiers and provide doctors with an explainability framework they can use for evaluating such models.

In this paper, we present the results of applying such models as well as their interpretations and confirm a few clinically known interactions. We start with a short overview of the dataset involved and the challenges faced when doing pre-processing. The ultimate usefulness of our results lies in the ability of the clinician to interpret and understand the main drivers behind the predictions. We take a look at the characteristics of the population of miss-classified predictions and we delve deeper into the model's choices by using the recent LIME explanation model described in Ribeiro et al. (2016).

1.1 RELATED WORK

Most applications of ML tools to cancer survival prediction to date focused on the most common types of cancer, in particular breast cancer. Two general meta studies are presented by Cruz &

Wishart (2006) and Kourou et al. (2015) where they analyze more than 1500 papers on the topic of ML and cancer. Most of these focus on cancer diagnosis while only 10% focus on prediction/prognosis which they further divided among susceptibility, recurrence and survival. Furthermore, only a smaller fraction combine clinical data with molecular data. The work of Park et al. (2010) creates a preoperative scale based on clinical factors to predict survival and closely resembles ours. We aim to have a comparison with their method for future work. Abreu et al. (2016) meta study is amongst the high quality studies that include relevant steps for applying ML models to clinical datasets, from data cleaning to the specifics of predicting and evaluating of breast cancer recurrence. For our models, we assess the performance using the accuracy and cross entropy loss, but the main focus remains interpretability.

1.2 Dataset

For this project we had access to anonymized patient information (observations) and in some cases several observations for the same subject. The dataset contained clinical features, some of which were age, gender, Karnofsky performance score (IK), tumor types and locations as well as well known molecular prognostic markers such as IDH1/2 mutations, MGMT, TERT, with a total number of 7630 observations, recorded as a mix of continuous and categorical variables. They were collected over several decades by the neuro-oncology department of a large University Hospital in Europe and is one of the largest, and most unique, databases of brain tumors. For this analysis we first performed clinically informed pre-processing which resulted in the final dataset containing 2086 unique observations. Firstly we subdivided the data based on the tumor types, Glioblastoma multiforme (GBM) and others (Oligo, Astrocytoma), as GBMs are known to lack predictive structure. We kept only the last observation for each patient and removed implausible outliers. One of the important steps was adding known interaction effects such as the IDH-TERT mutations, or calculating additional variables such as age at surgery, and simplifying the coding of some of the molecular features. Lastly, we deleted variables known by the clinicians to be non-informative.

Traditional techniques of dealing with missing data in medical studies, such as available case analysis, pairwise deletion, and single imputation are sub-optimal and lead to both loss of statistical power and increased bias in the results (Bell et al., 2014; Little et al., 2012). The dataset, in particular the molecular part, was very sparse. Several molecular indicators lacked data for approximately 50% of observations or even 83% for some observations, such as MGMT gene. Since complete case analysis was impossible, we needed a statistically robust method for handling the missing data. Using multivariate tests from Little et al. (2012) we confirmed the data is missing completely at random (MCAR). We kept variables with only 50% or more complete observations, and followed guidelines from Graham (2009) to impute 40 datasets using the MICE imputation method (Van Buuren, 2018), since this does not assume the data follows a multivariate normal distribution.

2 Methods and Results

The purpose of this analysis was to classify patients based on their survival after surgery and investigate what drives the model's predictions such that these tools can be used by clinicians. The first step was to construct a target variable by subtracting the surgery date from the observed death date (no censoring) or from the last visit (right censoring). Clinicians suggested we split our analysis in two datasets, glioblastoma patients only, and patients with all types of tumours.

2.1 Clustering

We briefly mention the results from two approaches for hierarchical clustering of the features. One of the methods that was based on PCAMIX (Kiers, 1991) aimed at making clusters homogeneous, with this homogeneity criteria proportionally increasing with the link between cluster variables. This method is able to confirm previously known clinical interactions, namely that TERT and IDH-TERT are highly correlated and drive the cluster formations. The second most important drivers are IDH1/2, tumour type and tumour grade.

Table 1: Experiment Results for **3 classes** for GBM/ALL Datasets

Model	**GBM/ALL Accuracy (%)**	**GBM/ALL LogLoss**
Logistic Regression	$0.47/0.60 \pm .02$	$0.94/0.84$
Random Forests	$0.57/0.61 \pm .02$	$0.94/0.85$
Neural Network Ensemble	$0.59/0.63 \pm .02$	$0.90/0.75$

The second one is COBWEB3 (MacLellan et al., 2016), a model that learns conceptual spaces around each cluster. COBWEB3 results in 2 large clusters and 6 smaller clusters. The resulting two main clusters exhibit a difference of a factor of 4 for average survival, with even higher differences as we narrow down on smaller clusters. Both clustering approaches provide the clinician with a tool to interpret and visualize the dataset and its statistics in an intuitive fashion, further allowing them to understand their patients in relation to other patients.

2.2 Predictive Modelling for survival prediction

For the purpose of this analysis we transformed the continuous target variable into a discrete variable, using 3 balanced classes. We use 400 days and 900 days as cut off points for the full dataset, and 300, 550 for the GBM dataset. We set aside 20% of the data for testing only. All results where averaged over 5 runs, as can be seen in Table 1.

The multinomial logistic regression model uses 5 fold cross-validation with l2 penalty and a low regularization coefficient of 0.01. We didn't find that a random forest model would perform much better. Through cross-validation and grid-search we found that the best criteria to split on is the *entropy* with a depth of 80 and a mix of 400 decision trees. Finally, for neural networks, our benchmark model was a 3-layer neural network, with one node per input variable, equivalent to 53 input nodes after one-hot-encoding of categorical variables[1]. We train our models between 100 and 150 epochs with a small batch-size of 32. Next, we probe for performance gains by widening and deepening the network structure. Furthermore, we add dropout between every layer in the network (including input layer), to add regularization. The best performing model is an ensemble of Neural Networks, with 4 dense layers and Relu activation functions. We train each one with a standard Adam optimizer and a learning rate of 0.001, and use a soft voting classifier to combine the results.

2.2.1 Interpretability

The ultimate purpose of this analysis was to provide value to the clinician and aid them in their decision-making. Their suggestion was to look at the entire dataset and dissect the model's choices, first by looking at the sub-population of miss-classified (false positives) predictions. This helped them understand whether the model cannot capture situations that were difficult for them as well or whether the model was not trustworthy. Our novel approach was to make use of the relatively new field of explainable AI and use the LIME method (Ribeiro et al., 2016) to understand the main features influencing our black box-predictions. The LIME explanation model works by perturbing a test instance locally (sampling in the vicinity of that instance) and building a linear prediction model. This in turn can be used to observe changes in predictions and allows for the associated weights to be used as interpretations. Furthermore, the method also allows one to test whether a model is globally truthful, by interpreting a carefully selected representative sample of the data (SP-LIME method). This was the better approach that eventually helped the clinician understand what drives predictions for individual patients but also whether the approach can be trusted.

The first step in analyzing our predictions was to look at the confusion matrix and observe the proportion of values for 3 main features of the 11 patients that we predicted to live more than 900 days when in fact they lived less than 400 days (right of Figure 1). We can see that the majority are men and they have tumor grade 3. By looking at a random instance from this small subset we observe what are the main drivers that influence the classifier into making the wrong decision. For

[1] One-hot-encoding was used for all categorical variables when modelling

example, we observe that the main drivers were the P16 gene being normal as well as high IK. The biospsy surgery type influences the classifier towards the right prediction.

Secondly, a clinician can and should ask whether this classifier is indicative of the truth or is in some way biased or wrong. Using the SP-LIME method we can select a number of representative examples that will give us a global view of the explanations. We perform sub-modular picking (SP) by selecting 10 representative explanations and highlight one of them in Figure 2. This particular example was useful for the clinician as he was able to understand to what extent features are important.

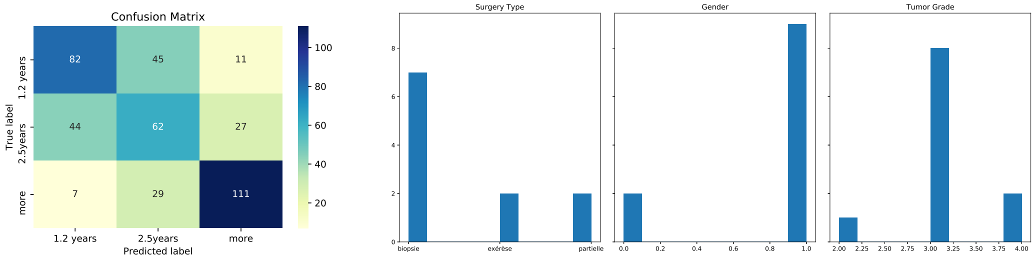
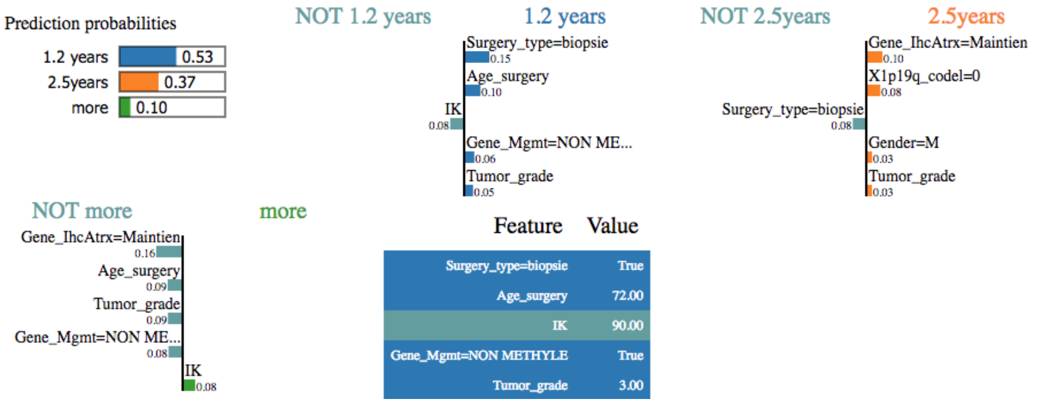


Figure 1: **Left**: Confusion Matrix for Random Forests **Right**: Histogram of 3 feature values for the 11 False Positives in the confusion matrix



Feature	Value
Surgery_type=biopsie	True
Age_surgery	72.00
IK	90.00
Gene_Mgmt=NON METHYLE	True
Tumor_grade	3.00

Figure 2: Example of a representative explanation constructed using SP-LIME. The patient was correctly classified to live less than 400 days. As one might expect for someone with a short lifespan, the main drivers are the age at surgery and surgery type as well as tumour grade, whereas the high performance score (IK) would pull the classifier in the opposite direction.

3 Discussion

This paper presented an initial analysis of a challenging and diverse clinical dataset involving brain tumour patients. Through preprocessing and applying hierarchical clustering methods, we provide doctors with a flexible framework, allowing them to visualize individuals and groups of patients as well as comparing their features. Secondly, we build supervised classification models that make a positive steps towards predicting patient survival, something clinicians expressed a particular interest in. Looking beyond the results, the focus was to add an interpretability layer on top of black box classifiers so that clinicians can assess the usefulness and reliability of such models, something we were able to confirm. Furthermore, we see interpretability play an important role in this analysis, that can be seen as a first step in a bigger pipeline that can potentially aid clinicians with their decision making. For future work, we plan to further interpret the clinical results that result from this analysis and improve the accuracy of the models as well as add a survival analysis model to compare it to regression models for survival. Finally, we envision enriching the dataset with images.

References

Pedro Henriques Abreu, Miriam Seoane Santos, Miguel Henriques Abreu, Bruno Andrade, and Daniel Castro Silva. Predicting breast cancer recurrence using machine learning techniques: a systematic review. *ACM Computing Surveys (CSUR)*, 49(3):52, 2016.

Zeynettin Akkus, Alfiia Galimzianova, Assaf Hoogi, Daniel L Rubin, and Bradley J Erickson. Deep learning for brain mri segmentation: state of the art and future directions. *Journal of digital imaging*, 30(4):449–459, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Melanie L Bell, Mallorie Fiero, Nicholas J Horton, and Chiu-Hsieh Hsu. Handling missing data in RCTs; a review of the top medical journals. *BMC medical research methodology*, 14:118, nov 2014. ISSN 1471-2288. doi: 10.1186/1471-2288-14-118. URL http://www.ncbi.nlm.nih.gov/pubmed/25407057http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4247714.

Joseph A Cruz and David S Wishart. Applications of machine learning in cancer prediction and prognosis. *Cancer informatics*, 2:117693510600200030, 2006.

O Gallego. Nonsurgical treatment of recurrent glioblastoma. *Current oncology*, 22(4):e273, 2015.

John W. Graham. Missing Data Analysis: Making It Work in the Real World. *Annual Review of Psychology*, 60(1):549–576, 2009. ISSN 0066-4308. doi: 10.1146/annurev.psych.58.110405.085530. URL http://www.annualreviews.org/doi/10.1146/annurev.psych.58.110405.085530.

HA Haenssle, C Fink, R Schneiderbauer, F Toberer, T Buhl, A Blum, A Kalloo, A Hassen, L Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 2018.

Henk AL Kiers. Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56(2):197–212, 1991.

Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

Jiangwei Lao, Yinsheng Chen, Zhi-Cheng Li, Qihua Li, Ji Zhang, Jing Liu, and Guangtao Zhai. A deep learning-based radiomics model for prediction of survival in glioblastoma multiforme. *Scientific reports*, 7(1):10353, 2017.

Roderick J. Little, Ralph D'Agostino, Michael L. Cohen, Kay Dickersin, Scott S. Emerson, John T. Farrar, Constantine Frangakis, Joseph W. Hogan, Geert Molenberghs, Susan A. Murphy, James D. Neaton, Andrea Rotnitzky, Daniel Scharfstein, Weichung J. Shih, Jay P. Siegel, and Hal Stern. The Prevention and Treatment of Missing Data in Clinical Trials. *New England Journal of Medicine*, 367(14):1355–1360, oct 2012. ISSN 0028-4793. doi: 10.1056/NEJMsr1203730. URL http://www.nejm.org/doi/abs/10.1056/NEJMsr1203730.

David N Louis, Arie Perry, Guido Reifenberger, Andreas Von Deimling, Dominique Figarella-Branger, Webster K Cavenee, Hiroko Ohgaki, Otmar D Wiestler, Paul Kleihues, and David W Ellison. The 2016 world health organization classification of tumors of the central nervous system: a summary. *Acta neuropathologica*, 131(6):803–820, 2016.

C.J. MacLellan, E. Harpstead, V. Aleven, and K.R. Koedinger. Trestle: A model of concept formation in structured domains. *Advances in Cognitive Systems*, 4, 2016.

V Panca and Z Rustam. Application of machine learning on brain cancer multiclass classification. In *AIP Conference Proceedings*, volume 1862, pp. 030133. AIP Publishing, 2017.

John K Park, Tiffany Hodges, Leopold Arko, Michael Shen, Donna Dello Iacono, Adrian McNabb, Nancy Olsen Bailey, Teri Nguyen Kreisl, Fabio M Iwamoto, Joohee Sul, et al. Scale to predict survival after surgery for recurrent glioblastoma multiforme. *Journal of clinical oncology*, 28(24): 3838, 2010.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. Why should i trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. ACM, 2016.

Stef Van Buuren. *Flexible imputation of missing data*. Chapman and Hall/CRC, 2018.