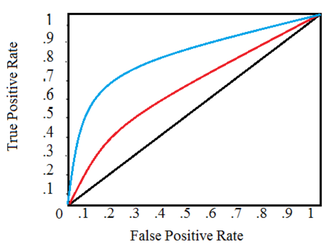
Machine learning prediction of glioblastoma patient one-year survival

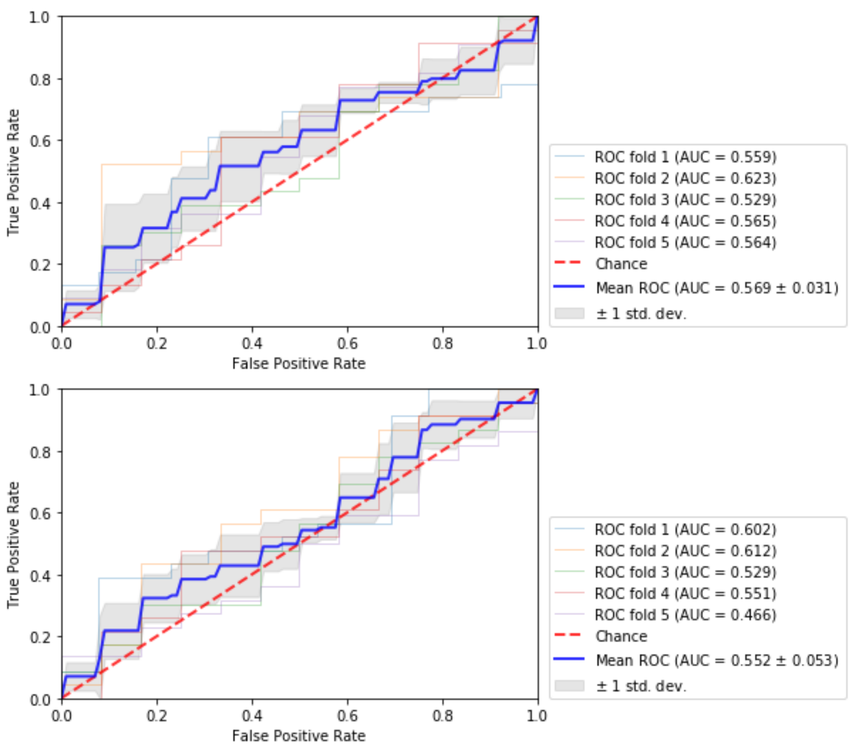
Andrew Du[1], Warren McGee[2], Jane Y. Wu[2]
[1]Illinois Mathematics and Science Academy, Aurora, Illinois, USA
[2]Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, Illinois, USA

Glioblastoma (GBM) is a grade IV astrocytoma formed primarily from cancerous astrocytes and sustained by intense angiogenesis. GBM often causes non-specific symptoms, creating difficulty for diagnosis. This study aimed to utilize machine learning techniques to provide an accurate one-year survival prognosis for GBM patients using clinical and genomic data from the Chinese Glioma Genome Atlas. Logistic regression (LR), support vector machines (SVM), random forest (RF), and ensemble models were used to identify and select predictors for GBM survival and to classify patients into those with an overall survival (OS) of less than one year and one year or greater. With regards to overall survival, a significant ($p < 0.05$, $n = 175$) correlation was found with age (negative), radiation treatment (positive), and chemotherapy treatment (positive). IDH1 mutation and 1p19q codeletion showed insignificant correlation with OS in this dataset. This potentially implies that IDH1 mutation alone, although important in secondary GBM prognosis, is insignificant for primary GBM prognosis. 1p19q codeletion also appeared to be insignificant for primary GBM prognosis when considered alone. The ensemble model had the highest overall accuracy, achieving a mean AUC score of 0.644 and an F1 score of 0.799.



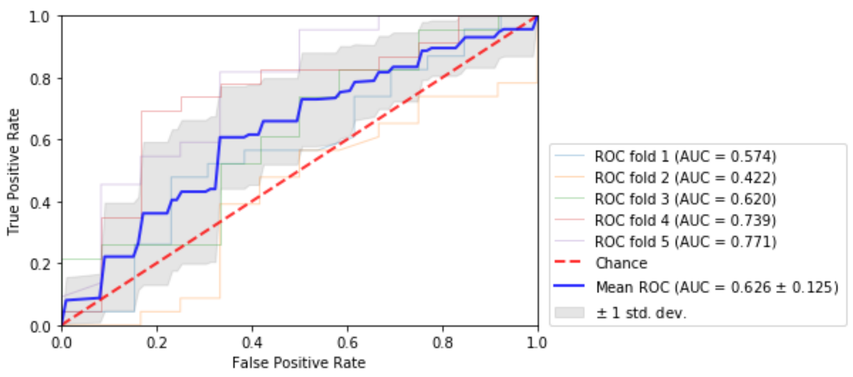
(left, image credit: Medium). *Sample ROC curve*. A receiver operating characteristic (ROC) curve "plots the true positive rate (TPR) versus the false positive rate (FPR) as a function of the model's threshold." "The threshold represents the value above which a data point is considered in the positive class." (Medium). In this study, survival of 365 days or greater is considered a "positive" classification while survival of less than 365 days is considered a "negative" classification. The area under the curve (AUC) quantifies the model's performance as a metric between 0 (worse) and 1 (better). In the figure, the blue curve has a higher AUC than the red curve and thus would be considered better performing.

LR, SVM classifiers tend to underfit the dataset



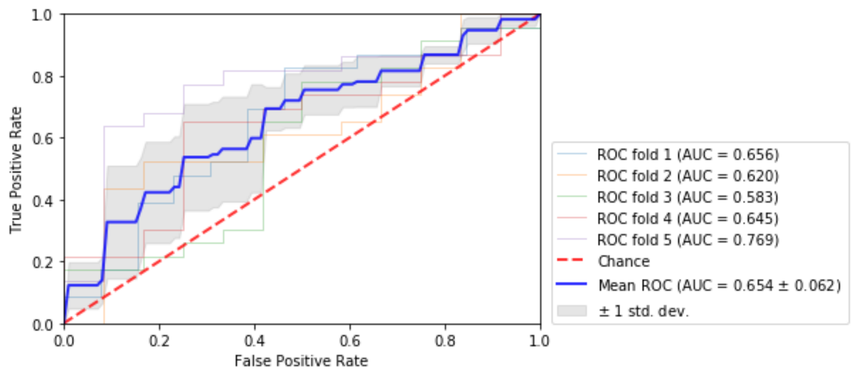
Underfitting indicates that the model is not specific enough to the training data.

RF classifier tends to overfit the dataset



Overfitting indicates that the model is too specific to the training data.

Underfitting of LR, SVM counteracts overfitting of RF in ensemble classifier



Figures 3, 4, 5, 6 (top to bottom). *Mean receiver operating characteristic curves*. Mean receiver operating characteristic curves were generated for LR (3). SVM (4), RF (5), and ensemble (6) models trained and tested using stratified 5-fold cross validation. AUC scores are shown in the figure legends.

OS correlation: Negative – age; Positive – radiation, chemo status

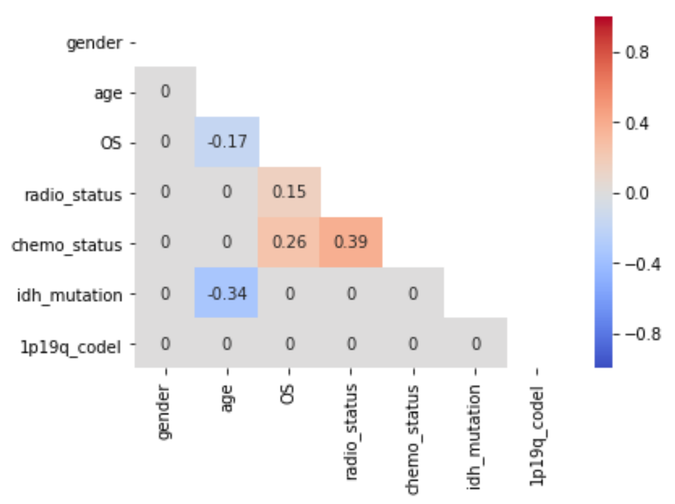


Figure 1 (left). *Correlation analysis of clinical factors*. Pearson correlation r values were determined for clinical factors in relation to each other. All statistically insignificant r values are shown as "0". Significant positive correlation was found between overall survival and radiation treatment status, between overall survival and chemotherapy treatment status, and between radiation treatment status and chemotherapy treatment status ($n = 176$, $r_{crit} = 0.148$, $p < 0.05$). Significant negative correlation was found between overall survival and age and between age and IDH1 mutation status ($n = 175$, $r_{crit} = -0.148$, $p < 0.05$).

Genomic factors are significantly correlated with OS

(Selected graphs shown)

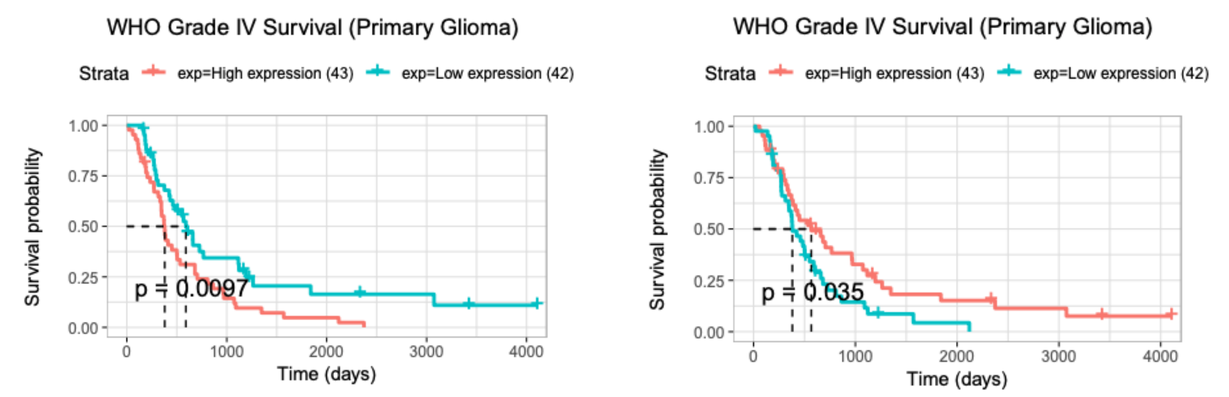


Figure 2 (above). *Survival probability vs time for high and low expression strata of DYX1C1-CCPG1 (above left) and RP11-355I22.2 (above right)*. Survival probability was plotted against time for patients in equally sized strata of high and low expression of the respective genes. DYX1C1-CCPG1 (above left) was negatively correlated with overall survival (r = –0.297), and RP11-355I22.2 (above right) was positively correlated with overall survival (r = 0.450) in the selected cohort. Time at which survival probability = 0.50 was analyzed for statistical significance (p-values shown on graphs).

Ensemble classifier had the highest accuracy

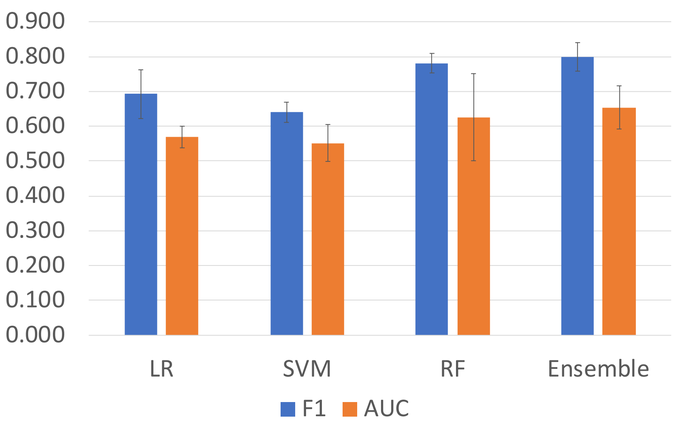


Figure 7 (left). *F1 and AUC scores for each classifier*. F1 (harmonic mean of precision and recall; see **below**, credit: Medium) and AUC scores for each classifier were calculated. Population standard deviation was calculated for each value among the five stratified folds for each classifier.

$$recall = \frac{true\ positives}{true\ positives + false\ negatives} \qquad precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

SUMMARY

Age was negatively correlated with overall survival, while radiotherapy and chemotherapy status were positively correlated with overall survival. The ensemble classifier exhibited the highest accuracy compared to the LR, SVM, and RF classifiers alone – the LR and SVM classifiers' underfitting tendency appeared to counteract the RF classifier's overfitting tendency.

Acknowledgements

Dr. Jane Y. Wu and Dr. Warren McGee provided significant support and guidance in this study. The IMSA SIR department provided transportation to the research site.

References

Tamimi AF, Juweid M. Epidemiology and Outcome of Glioblastoma. In: De Vleeschouwer S, editor. Glioblastoma [Internet]. Brisbane (AU): Codon Publications; 2017 Sep 27. Chapter 8. Available from: https://www.ncbi.nlm.nih.gov/books/NBK470003/doi: 10.15586/codon.glioblastoma.2017.ch8
Stupp R, Mason WP, van den Bent MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med. 2005;352(10):987-996.
Blomqvist P, Lycke J, Strang P, Törnqvist H, Ekbom A. Brain tumours in Sweden 1996: care and costs. J Neurol Neurosurg Psychiatry. 2000;69(6):792–798. doi:10.1136/jnnp.69.6.792
Zuo S, Zhang X, Wang L. A RNA sequencing-based six-gene signature for survival prediction in patients with glioblastoma. Sci Rep. 2019;9(1):2615. Published 2019 Feb 22. doi:10.1038/s41598-019-39273-4
Jia D, Li S, Li D, Xue H, Yang D, Liu Y. Mining TCGA database for genes of prognostic value in glioblastoma microenvironment. Aging (Albany NY). 2018;10(4):592–605. doi:10.18632/aging.10141
Weller, M., Stupp, R., Reifenberger, G. et al. MGMT promoter methylation in malignant gliomas: ready for personalized medicine?. Nat Rev Neurol 6, 39–51 (2010) doi:10.1038/nrneurol.2009.197
Lalezari S, Chou AP, Tran A, et al. Combined analysis of O6-methylguanine-DNA methyltransferase protein expression and promoter methylation provides optimized prognostication of glioblastoma outcome. Neuro Oncol. 2013;15(3):370–381. doi:10.1093/neuonc/nos308
Leili Shahriyari, Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma, Briefings in Bioinformatics, Volume 20, Issue 3, May 2019, Pages 985–994, https://doi.org/10.1093/bib/bbx153
Macyszyn L, Akbari H, Pisapia JM, et al. Imaging patterns predict patient survival and molecular subtype in glioblastoma via machine learning techniques. Neuro Oncol. 2016;18(3):417–425. doi:10.1093/neuonc/nov127
Lv S, Teugels E, Sadones J, et al. Correlation between IDH1 gene mutation status and survival of patients treated for recurrent glioma. Anticancer Res. 2011;31(12):4457-4463.