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.

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 ($n = 175$, $r_{p} = 0.148$, $p < 0.05$). Significant negative correlation was found between overall survival and age and between age and IDH1 mutation status ($n = 175$, $r_{p} = -0.148$, $p < 0.05$).

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

#### Genomic factors are significantly correlated with OS

### Ensemble classifier had the highest accuracy

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

---

**References**


**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.

**Figures**

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.