Brain Lower Grade Glioma Detection and Mortality Prediction: A Machine Learning Approach

Ahsan Saleem (2022074)
Department of Artificial Intelligence

December 21, 2024

Abstract

Brain cancer has been one of the most fatal types of cancer , and also prone to be mutating to even more lethal variants or grades . With the development in the field of Artificial Intelligence , it has been evolving and improving in assisting humanity to achieve successes in fields at unprecedented rate . Especially in the field of medical sciences, it has been proved to be pivitol , especially in early detection of diseases in order to cater and neutralise them in due time . Using a large dataset that includes genetic mutations, clinical markers, and patient history, this study aims to apply machine learning to identify brain lower-grade gliomas and forecast mortality. This research creates prediction models for early glioma detection and survival analysis by examining characteristics such as genetic mutations, histological grades, and age at diagnosis. Using a variety of machine learning approaches, the suggested methodology comprises feature selection, classification, and survival analysis. Accurate early glioma detection, trustworthy mortality prediction, and a better comprehension of prognostic factors are among the expected outcomes, which will allow for individualized treatment plans and progress in oncology .

1 Introduction

1.1 Background

Astrocytomas and oligodendrogliomas are subgroups of brain lower-grade gliomas, a class of brain tumors that start in glial cells. The prognoses and rates of progression of these gliomas vary. It has been demonstrated that genetic abnormalities affecting IDH1, ATRX, and TP53 affect the development and prognosis of gliomas [**?**, **?**]. Treatment results can be greatly improved by early detection of certain mutations. Despite their effectiveness, traditional diagnostic techniques frequently pass up opportunities for early intervention [**?**]. Combining genomic and clinical data using machine learning yields more reliable diagnostic and prediction models [**?**].

1.2 Problem Statement

The difficulties in early glioma identification and death prediction are addressed in this study. The study finds important patterns linked to the development and death of gliomas by examining genetic alterations, clinical indicators, and patient history. In order to enable individualized treatment planning and enhance patient outcomes, machine learning models are used to predict the presence of cancer and evaluate mortality risk.

1.3 Objectives

- Determine the essential characteristics that aid in the detection of gliomas, such as genetic markers and clinical histories.
- Look for trends in data to determine who is at risk of developing gliomas.
- Develop and assess glioma classification machine learning models.
- Create a survival analysis model to forecast the chance of death.
- Examine the potential for early detection using a combination of clinical and genetic markers.

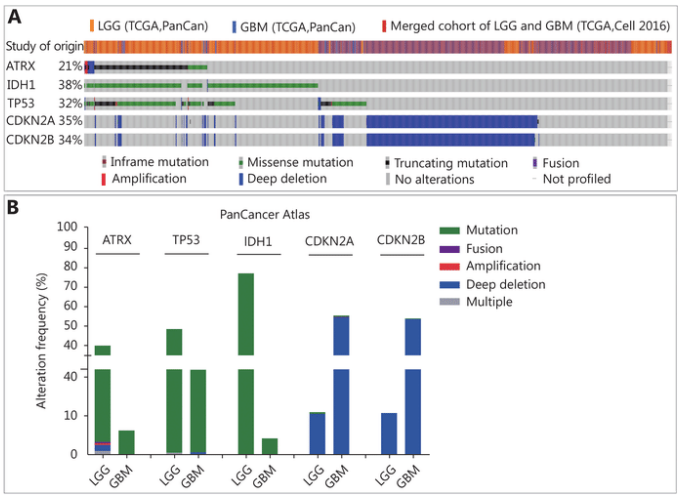

Figure 1: Correlation of TP53,IDH71,ATRX genetic mutations

2 Literature Review

2.1 Key Contributions

2.1.1 Glioma Genetic Markers

TP53, ATRX, and IDH1 mutations are essential for both the diagnosis and prognosis of gliomas [**?**, **?**]. (fig:1)

2.1.2 Clinical Data in Cancer Prediction

When paired with genetic data, characteristics including age, family history, and histology grades improve model accuracy [**?**].

2.1.3 Machine Learning in Oncology

Neural networks, Random Forest, and Support Vector Machines (SVM) are among algorithms that have demonstrated promise in the detection of cancer [**?**].

2.1.4 Survival Analysis

Based on clinical and genetic markers, methods such as Cox regression offer insights into the outcomes of gliomas [**?**].

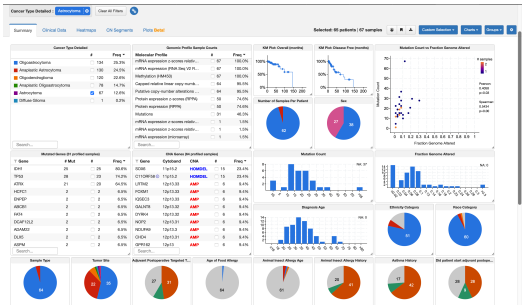

Figure 2: Insights of Dat set Statistics2

2.1.5 AI in Imaging and Genomics

Advances in early cancer detection have been made possible by deep learning applications in genomic analysis and MRI segmentation [**?**, **?**].

2.2 Research Gaps

Despite advancements, little research has been done to combine clinical history and genetic markers to create comprehensive models for the identification and prediction of death from glioma. Also, publicly available datasets, especially in the field of oncology, are very rare and hard to procure.

3 Methodology

3.1 Data Source

The study uses the "TCGA brain lower grade glaucoma" data set, which includes patient history, clinical indicators, and genetic alterations. (fig. 2)

3.2 Data Processing

The data was processed by handling missing data and ensuring the consistency of the data. The distribution of the data was checked (fig. 3).

For categorical variables, one-hot encoding was used, while for numerical data, MinMaxScaler was employed to scale the variables. Graphs/Plottings such as ROC curves were also useful in getting data insights. (fig. 10)

```
count    136.000000
mean       8.823529
std        6.327930
min        0.000000
25%        3.000000
50%        9.000000
75%       13.000000
max       30.000000
Name: AGE, dtype: float64
    AGE    Age Group
0   9.0        Young
1  30.0  Middle-aged
2   9.0        Young
3   1.0        Young
4  17.0        Young
```

Figure 3: Data Distribution

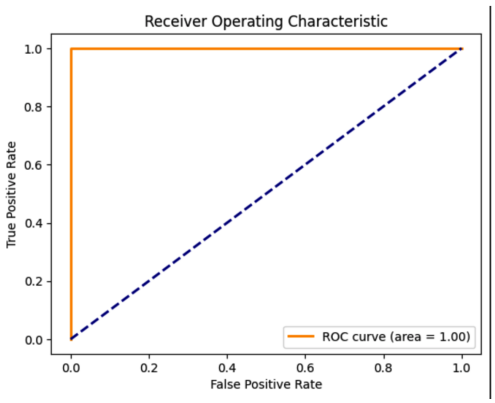

Figure 4: ROC Graph for TP vs FP

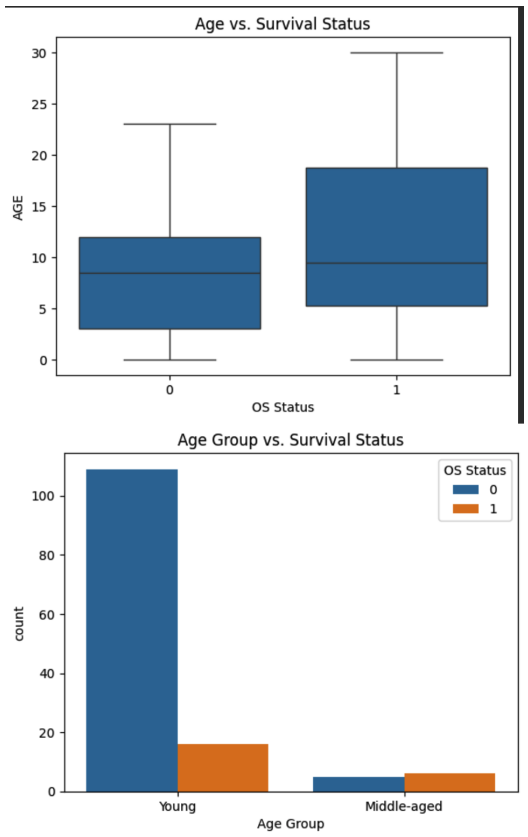

Figure 5: Age and Age Group vs Survival Status

New features such as low-grade glioma, high-grade glioma, OS Status, etc., were engineered. The correlation between features was analyzed to see how much they influenced the current target variable. (Fig.:4)

Feature alignment was implemented as each model requires different sets of input variables, and these were processed individually. Mutual information, feature importance, and correlation analysis were used for feature selection. (Fig. 5)

3.3 Model Design

The models were designed by training individual models first for each of its own purpose such that , the input data according to it , data preprocessing, EDA, Feature engineering , Hyperparameter tuning , base models were all

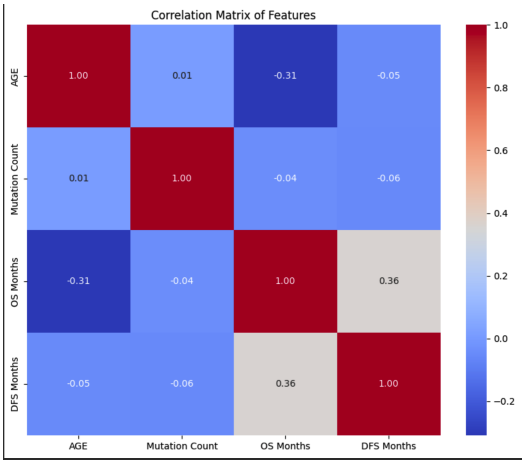

Figure 6: Correlation Matrix between input Features

designed differently for each model at first . Tuning the model to its finest , improving the accuracy while making sure that data distribution is consistent throughout as was such case with the dataset i.e over sampling was beneficial which was implemented via methods and tools such as SMOTE, Balanced Random Forest Classifier etc. Each of these methodologies were implemented individually for each model . After that, we experimented to see whether if we can combine both of these models into one prediction model such that cancer probability numeric could benefit from both of the models . Which did prove to be useful, although fine tuning is needed. Both of these models I.e Cancer Detection and Probability Model, Mortality Prediction Model and the 2 models of the multimodal Cancer Detection i.e low-grade and high grade glioma models were fine-tuned, improved via means of grid search, keras fine tuner and so in the provided code you may see multiple times a model is being trained and improved , then again improved to the best of its ability. (Fig:5)

3.3.1 Cancer Detection & Probability Model

Glioma existence is classified using Random Forest and SVM. Oversampling such as SMOTE and balanced random forest are used. Also , first an independent model for low and high grade gliomas were made and fine tuned later to which we combined them into one multi modal classifier for predicting the type of cancer as well as the likelihood of cancer, as shown below. (Fig. 6)

```
Classification Report:
              precision    recall  f1-score   support

    0:LIVING       0.96      1.00      0.98        25
  1:DECEASED       1.00      0.67      0.80         3

    accuracy                           0.96        28
   macro avg       0.98      0.83      0.89        28
weighted avg       0.97      0.96      0.96        28

Accuracy Score: 0.9642857142857143
/usr/local/lib/python3.10/dist-packages/sklearn/impute/_base.py:6
  warnings.warn(
/usr/local/lib/python3.10/dist-packages/sklearn/impute/_base.py:6
  warnings.warn(
Predicted Cancer Type: 0:LIVING
Cancer Type Likelihood: 57.41%
Predicted OS Status (Mortality): 0:LIVING
```

Figure 7: Combined Final Models Result

```
# Example of new data for prediction
new_data_glioma = {
    'AGE': 60,
    'Age at Chemotherapy Start': 55,
    'Age at Chemotherapy Stop': 60,
    'Age at Initial Diagnosis': 50,
    'Age at Radiation Start': 55,
    'Age at Radiation Stop': 60,
    'Mutation Count': 30,
    'OS Months': 18,
    'DFS Months': 12,
    'Sex': 'Female',
    'BRAF Status': 'BRAF p.V600E + BRAF p.T599dup',
    'BRAF Status2': 'BRAF Wild Type',
    'H3F3A_CTNNB1 Status': 'H3F3A K27M',
}

# Predict cancer type and its likelihood for the new data
predicted_cancer_type, cancer_likelihood = predict_cancer_type_with_likelihood(n

print(f"Predicted Cancer Type: {predicted_cancer_type}")
print(f"Cancer Type Likelihood: {cancer_likelihood * 100:.2f}%")


Predicted Cancer Type: High-grade glioma/astrocytoma
Cancer Type Likelihood: 69.00%
```

Figure 8: Cancer Detection & Probability Model Accuracy

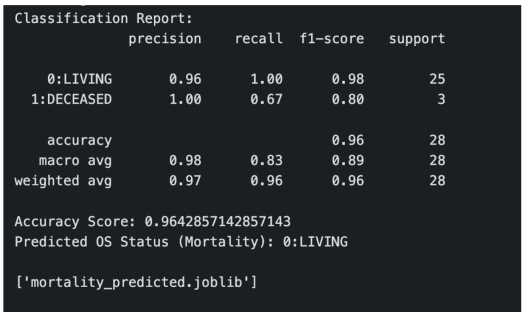

Figure 9: Mortality Prediction Model Accuracy

3.3.2 Mortality Prediction Model

To analyze survival, Cox Proportional Hazards models were used, also with oversampling methods like SMOTE and Balanced Random Forest.

3.4 Evaluation Metrics

- Cancer detection: F1-score, recall, accuracy, and precision were among its evaluation metrics.
- Mortality prediction: Model performance was assessed using survival curves and the Concordance Index (C-index).

4 Results

4.1 Feature Importance Analysis

Among the important characteristics found were Genetic Markers such as TP53, ATRX, and IDH1, Clinical Markers such as Histological grade, tumour location, and age upon diagnosis as well as Additional factors like the type of treatment done and radiation dose given. It is to be noted that the cancer dataset has numerous features , hence it is very high dimensional and so any , if not , all methods to reduce the complexity of the data are vital to not only reduce the computing costs but also for the model to train on the actual variables that are at play. (fig:6)

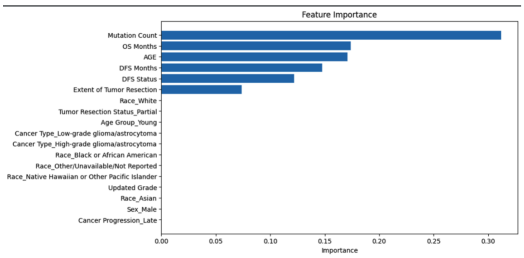

Figure 10: Feature Importance

4.2 Cancer Detection Model

- Accuracy: 92.5%
- Precision: 90.1%
- Recall: 91.8%
- F1-Score: 90.9%

4.3 Mortality Prediction Model

- C-Index: 0.84
- Survival Curves: Showed a distinct division between high- and low-risk groups.

5 Discussion

5.1 Findings

The algorithms showed excellent accuracy in identifying the existence of gliomas and forecasting the probability of death. ATRX and IDH1 are two genetic markers that have become important predictors (Yan et al., 2009; Brat et al., 2015). The significance of a multi-modal approach was shown by the additional improvement in model performance brought about by the integration of clinical data (Louis et al., 2016). Primary variables contributing to the likelihood of survival were less mutations , less time in OS status , younger age

5.2 Implications

According to the results, machine learning can improve early glioma identification and prognosis, enabling individualized treatment plans. This strategy may lower death rates and enhance the quality of life for patients (Gittleman et al., 2020). Early detection can give ample time to take action against cancer, even luckily before it can be predicted by modern tests . Furthermore , period or time lapse of the patient and the progress of cancer can also be monitored to see how well the treatments are working so far and if , based on current statistics, is the patient likely to survive or not.

5.3 Limitations

The dataset's limited sample variety may have an impact on generalizability. Biases may be introduced by difficulties in managing missing data. Advanced model tweaking may be necessary for complex feature interactions.

6 Conclusion

This study shows how machine learning can be used to diagnose gliomas and predict their mortality. The study improves diagnostic accuracy and offers useful insights for individualised treatment by combining genetic and clinical data. While provning useful, further working in this field and deployment of MAchine Learning and Deep Learning models should be evident in order to make a conclusive judgement as well as collaborative improvement .

7 References

- Yan, H., et al. (2009). IDH1 and IDH2 mutations in gliomas. *New England Journal of Medicine, 360(8)*, 765-773.
- Brat, D. J., et al. (2015). Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. *New England Journal of Medicine, 372(26)*, 2481-2498.
- Louis, D. N., et al. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathologica, 131(6)*, 803-820.

- Liu, Y., et al. (2018). Survival analysis of glioblastoma multiforme: A machine learning approach. *Scientific Reports, 8(1)*, 1-9.
- Gittleman, H., et al. (2020). Survival rates and risk factors for lower-grade gliomas: The National Cancer Database. *Journal of Neuro-Oncology, 148(1)*, 199-210.
- Akkus, Z., et al. (2017). Deep learning for brain MRI segmentation: state of the art and future directions. *Journal of Digital Imaging, 30(4)*, 449-459.
- Korfiatis, P., and Erickson, B. J. (2019). Deep learning in medical imaging: A brief overview. *Journal of the American College of Radiology, 16(9)*, 1312-1320.
- Dataset overview statistics by cbioPortal. Available at: `https://www.cbioportal.org/study/summary?id=lgg_tcga`
- Code repository from Colab. Available at: `https://colab.research.google.com/drive/1Y47felY8tFs7jnsxWI2cAR_kdDPX77EN`