Development of machine learning method for predicting patient prognosis in pan-cancer

Supervisor: Prof. WANG, Jiguang / LIFS

Student: LEUNG, Kung Nam / BTGBM Course: UROP1000/Summer

1. Abstract

With the lifetime risk of cancer uprising worldwide (Ahmad, A. S., Ormiston-smith, N. & Sasieni, P. D., 2015), cancer prognosis, or the evaluation of cancer development and survival analysis, is playing a more significant role in making treatment decisions. While clinical factors for prognosis include age, race, ethnicity, tumor histological type, and therapeutic history, this project aims to access the feasibility of incorporating clinical features alongside pan-cancer features, which are the data on frequently mutated genes and other genomic abnormalities (Illumina, Inc., 2022), for survival time prediction using machine learning methods. The project focuses on Glioblastoma multiforme (GBM) and carries out a fundamental analysis of the provided GBM dataset, followed by an evaluation of different regression techniques in survival prediction. The goal is to provide a foundation for further development of more advanced models on the dataset, including deep learning models.

2. Introduction

2.1 Project Objective

Given there is limited research on regression model for the survival estimation on the targeted dataset, the project aims to 1) Provide a proof-of-concept study on survival prediction (*Overall_Survival*) using machine learning methods on the specified dataset, and 2) Provide a coarse benchmark for evaluation of future advance models.

2.2 Glioblastoma multiforme

Glioblastoma multiforme (GBM), i.e. grade IV astrocytoma, is a fast-growing and aggressive brain tumor (American Association of Neurological Surgeons, 2022). It has an age-adjusted incidence rate of 3.22 per 100,000 population in the US (Chen, B. & Chen, C., 2021), with a five-year survival rate of 6.8 percent and an average length of survival of 8 months (National Brain Tumor Society, 2022).

2.3 GBM Dataset

The dataset of interest is retrieved from Broad Institute GDAC Firehose (Broad Institute of Mit & Harvard, 2016) under The Cancer Genome Atlas Program (TCGA), which collects pan-cancer clinical data spanning 33 cancer types since 2006 (National Cancer Institute, 2021).

Particularly, it consists of 660 features across 595 GBM patients, which 12 of them are clinical features (CLIs) and 648 of them are Genomic (pan-cancer) features. Out of the 595 patients, 490 of which are deceased and 105 of which are surviving.

Meanwhile, the genomic features are very sparse (appendix. 1), having over 70% of features being at least half-null. Note that the average survival days of patients is 504.5, which is about 2 times longer than 8 months from the existing statistics, implying a possible selection bias in TCGA.

3. Methodology

The general pipeline is illustrated in fig.1:

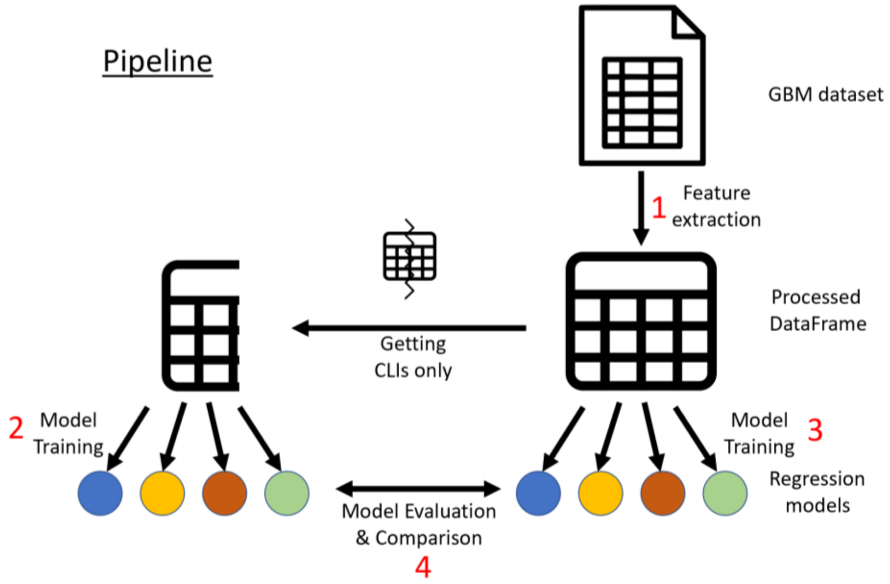


fig. 1 General Pipeline

3.1 Data Processing and Feature Extraction

Firstly, One-hot encoding is implemented to categorical features including *CLI_gender*, *CLI_radiation_therapy*, *CLI_histological_type*, *CLI_race,* and *CLI_ethnicity*. Note that to handle null values, parameter *drop_first* is set to false as the one-hot columns are not exhaustive, which zero in all columns represents previous null values.

Secondly, rows of surviving(non-deceasing) patients are dropped as they have not fully experienced their survival, and their survival days underestimate *Overall_Survival,* which cannot be used for training.

Thirdly, CLI_days_to_death is renamed to Overall_Survival and considered as the label.

Fourthly, null values in CLI_karnofsky_performance_score are replaced by the median.

Fifthly, null values in Genomic features are filled by zeros.

Correlation between *Overall_Survival* and CLIs is also studied (appendix 2). However, CLIs with low correlation are not dropped due to the possible intricate relationship to be explored in non-linear models e.g. Neural Network.

(Note: #CLIs increases from 12 to 14 after processing)

3.2 Model training for CLIs

As the number of features (660) is far greater than the number of instances (490) and in order to view the marginal effect of adding Genomic features, the 14 CLIs are selected from the dataset to first perform regressions.

Four regression models of varying complexity were chosen for the task, including 1) Linear Regression (and with Lasso regularization), 2) Random Forest Regression, 3) Single-layer Perceptron, and 4) Neural Network.

Train-test split(test_size=0.2) and normalization are applied on the processed dataset. To generate a benchmark model, the loss function Root Mean Squared Error (RMSE) is minimized and test set RMSE is used for model evaluation and comparison. Simple hyperparameter tunings are performed. Significant features are also expected to be highlighted by Lasso Regression for future modelling uses.

3.3 Model training for all features

The same modelling pipeline in 3.2 is repeated on the full dataset that contains both CLIs and Genomic features.

3.4 Model Evaluation & Comparison

All 8 models are evaluated based on their test set RMSE. While the four regressor types competes, the same regressor with or without appending Genomic features are also compared to view the marginal effect of Genomic features on survival prediction.

4. Results and Discussion

4.1.1 Linear Regression

	CLIs only	CLIs + Genomic features
Test RMSE	457.68	1047024939421714.0

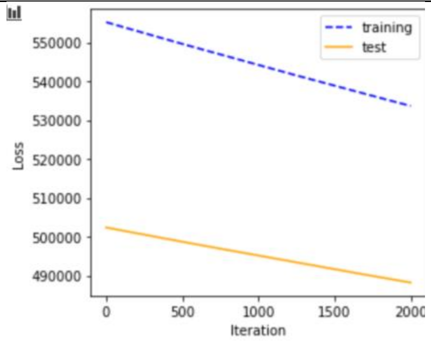
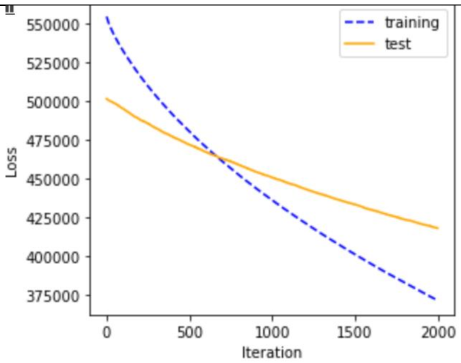
4.1.2 Linear Regression with Lasso Regularization

	CLIs only	CLIs + Genomic features
Test RMSE	**456.04(benchmark)**	471.32
alpha	20	55
#non-zero	9	6(CLIs)+5(Genomic features)
non-zero features	CLI_years_to_birth, CLI_date_of_initial_pathologic_diagnosis, CLI_karnofsky_performance_score, CLI_gender_male, CLI_radiation_therapy_no, CLI_radiation_therapy_yes, CLI_histological_type_treated primary gbm, CLI_ethnicity_hispanic or latino, CLI_ethnicity_not hispanic or latino	CLI_years_to_birth, CLI_date_of_initial_pathologic_diagnosis, CLI_karnofsky_performance_score, CLI_radiation_therapy_no, CLI_radiation_therapy_yes, CLI_histological_type_treated primary gbm, SMG_mutsig2.0_SFT2D1_cosmic, Amp_8q24.21, Del_1p36.32, Del_10p11.23, CN_1p_Amp

4.2 Random Forest Regression

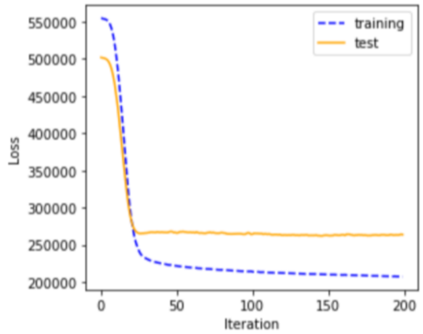
	CLIs only	CLIs + Genomic features
Test RMSE	531.99	539.95
n_estimators	65	65

4.3 Single-layer Perceptron

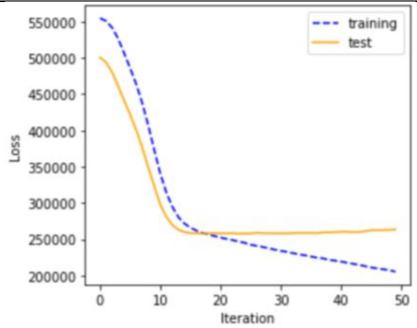
	CLIs only	**CLIs + Genomic features**
Test RMSE	698.73	646.81
batch_size	100	10
	 (still underfits at 2000 epochs)	 (still underfits at 2000 epochs)

4.4 Neural Network

<table>
<tr><th></th><th>CLIs only</th><th><b>CLIs + Genomic features</b></th></tr>
<tr><td>Test RMSE</td><td>513.55</td><td>503.89</td></tr>
<tr><td>batch_size</td><td>10</td><td>10</td></tr>
<tr><td>#hidden_layer</td><td>2</td><td>2</td></tr>
<tr><td>#node_first_layer</td><td>16</td><td>9</td></tr>
<tr><td>#node_second_layer</td><td>8</td><td>11</td></tr>
<tr><td></td><td>



</td><td>



</td></tr>
</table>

4.5 Result Summary

1. Out of all the models, Lasso Regression (alpha = 20) on CLIs only has minimal test set RMSE(456.04), thus it is the benchmark for further modelling.

2. In both groups(with or without appending Genomic Features), Lasso Regression has the lowest test set RMSE within the group.

3. Only Perceptron and Neural Network have improved test RMSE after incorporation of Genomic Features.

4.6 Discussions

- Given 4.5.3, since perceptron still underfits after 2000 epochs, Neural Network is the only viable option if Genomic features are the future interest of research and modelling.

- The extremely high RMSE for Linear Regression(without Lasso)(with Genomic Features) indicates the need for Regularization when modeling with Genomic Features.
- The 5 nonzero Genomic Features in Lasso Regression can become a natural selector of Genomic Features in future data preprocessing and modelling. Follow-up biochemical research on the relationship between these gene mutations and GBM is encouraged.
- The models(especially Neural Network) are unstable as RMSE and Loss/Iteration Curve vary vigorously after changes in random_states.

5. Recommendations

- Possible selection bias discussed in 2.3 is encouraged to be investigated
- The assumption to set null values in Genomic Features to 0 is arguable as some mutations can have negative values. Experimenting with more null-filling techniques are encouraged after throurough understanding on the Genomic Features.
- CNN can be tried in further studies, as it is intuitive to use convolution to test the pairwise functionality of mutations.
- P(number of features)(660)>N(number of samples)(490) could be a reason for an unstable model. Given the sparsity of the dataset, feature selection on Genomic Features and dimension reduction techniques e.g. PCA, SVD, and matrix factorization are encouraged in future studies to reduce the number of features.

6. Conclusion

To conclude, regression techniques are feasible in GBM survival prediction with Neural Network recommended if Genomic Features are the interest of research. A benchmark of RMSE=456.04 from Lasso Regression is provided as a foundation for the development of more advanced models in the future. The GBM dataset and source code for the project are available on GitHub (see Reference).

7. References

Source Code and Dataset https://github.com/Chowjai/2122Summer-UROP1000

Ahmad, A. S., Ormiston-smith, N. & Sasieni, P. D. (2015). Trends in the lifetime risk of developing cancer in Great Britain: comparison of risk for those born from 1930 to 1960. https://www.nature.com/articles/bjc2014606

Illumina, Inc. (2022). Pan-Cancer Analysis | Tumor type-agnostic studies using NGS panels. https://sapac.illumina.com/areas-of-interest/cancer/clinical-cancer-research/somatic-mutations/pan-cancer.html

American Association of Neurological Surgeons (2022). Glioblastoma Multiforme – Symptoms, Diagnosis and Treatment Options. https://www.aans.org/en/Patients/Neurosurgical-Conditions-and-Treatments/Glioblastoma-Multiforme

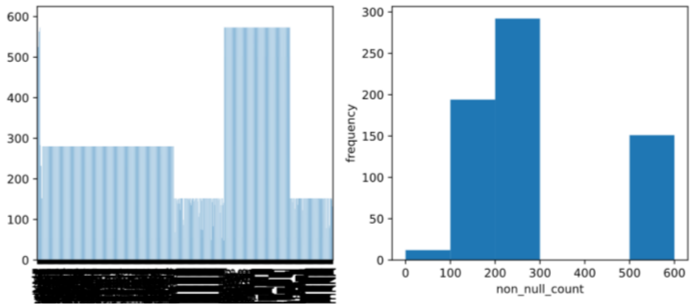
Chen, B. & Chen, C. (2021). Recent incidence trend of elderly patients with glioblastoma in the United States, 2000–2017. https://bmccancer.biomedcentral.com/articles/10.1186/s12885-020-07778-1#:~:text=etiology%20of%20glioblastoma.-,Background,malignant%20brain%20tumors%20%5B1%5D.

National Brain Tumor Society (2022). About GBM - GBM Awareness Day | National Brain Tumor Society. https://braintumor.org/take-action/about-gbm/

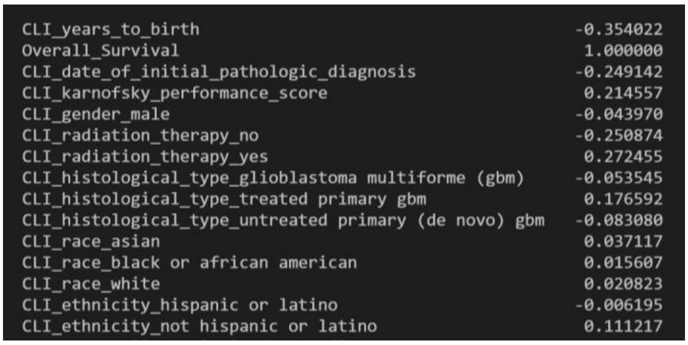
Broad Institute of Mit & Harvard (2016). Broad GDAC Firehose. https://gdac.broadinstitute.org/#

National Cancer Institute (2021). The Cancer Genome Atlas Program - NCI. https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

8. Appendix



appendix. 1 Sparsity of Genomic Features (number of non-null values for each mutation(left), number of Genomic Feature in different ranges of non_null_count(right))



```
CLI_years_to_birth                                    -0.354022
Overall_Survival                                       1.000000
CLI_date_of_initial_pathologic_diagnosis              -0.249142
CLI_karnofsky_performance_score                        0.214557
CLI_gender_male                                       -0.043970
CLI_radiation_therapy_no                              -0.250874
CLI_radiation_therapy_yes                              0.272455
CLI_histological_type_glioblastoma multiforme (gbm)   -0.053545
CLI_histological_type_treated primary gbm              0.176592
CLI_histological_type_untreated primary (de novo) gbm -0.083080
CLI_race_asian                                         0.037117
CLI_race_black or african american                     0.015607
CLI_race_white                                         0.020823
CLI_ethnicity_hispanic or latino                      -0.006195
CLI_ethnicity_not hispanic or latino                   0.111217
```

appendix. 2 Correlation between each CLIs and Overall_Survival

9. Glossary

GBM	Glioblastoma multiforme
CLIs	Clinical Features
RMSE	Root Mean Squared Error
TCGA	The Cancer Genome Atlas Program
Overall_Survival	Number of surviving days (the prediction label)