CSE 258 Assignment 2

Zhaoyang Jia
Department of Computer Science
UC San Diego La Jolla, CA 92093
z4jia@ucsd.edu

Gino Prasad
Department of Computer Science
UC San Diego La Jolla, CA 92093
giprasad@ucsd.edu

Task 1: Motivation, Explanation of Dataset and Exploratory Analysis

Motivation

Our project focuses on prediction of RNA expression of cancer genes. For the lay reader we will define some relevant terms:

Glossary:

DNA: In the nucleus every human cell, the DNA contains the genetic information on how to encode proteins. The vast majority of DNA is regulatory, meaning it doesn't directly translate into proteins but instead determines how much / how little of a protein is made.

RNA: The protein-coding sections of DNA are first converted into RNA sequences in a process known as transcription. These RNA sequences are converted to newly created proteins in a process known as translation. RNA can be thought of as an intermediary molecule between DNA and Proteins.

RNA-seq data: RNA-seq is represented by a vector of how many RNA sequences were contained in the cell for each gene. This is important because biologists use RNA-seq data to estimate the relative amount of a given protein in a sample (i.e., how many of a particular protein are in the sample).

Dataset Our dataset consists of RNA-seq data from *The Cancer Genome Atlas* [1], a publicly accessible repository of sequencing data from cancer patient tissue. Specifically, our dataset focuses on 516 *Lower Grade Glioma* tumor samples (brain cancer). This dataset consists **RNA-seq** data (Size 60660 vector gene expression counts) from the patient's tumor and a partial list of if and when patients have passed away. All samples from this dataset are publicly accessible and none of our below analysis is confidential.

Motivation: Our goal is to determine which genes are in the same **regulatory network** as known cancer-causing genes.

Exploratory Data Analysis: To better understand our dataset, we wanted to validate our analysis using the existing literature on cancer-promoting genes. Specifically, we surveyed the cancer genomics literature [2, 3, 4, 5, 6, 7] and found 7 genes that are known to be tumor-promoting in cancer research: MGMT, IDH1, IDH2, CIC, TP53, MKI67, MIB1. Namely, we want to confirm that these genes' over or under-expression (more or less RNA sequences found) correlates to a higher mortality rate (lower survivability).

To answer this question, we obtained the publicly available mortality information for our dataset, which contains a partial list of which patients had passed away, and how many days after RNA-sequencing. To simplify our exploratory analysis, we excluded individuals who did not have a recorded date of passing. If our analysis were to match the existing literature, we would that see

Preprint. Under review.

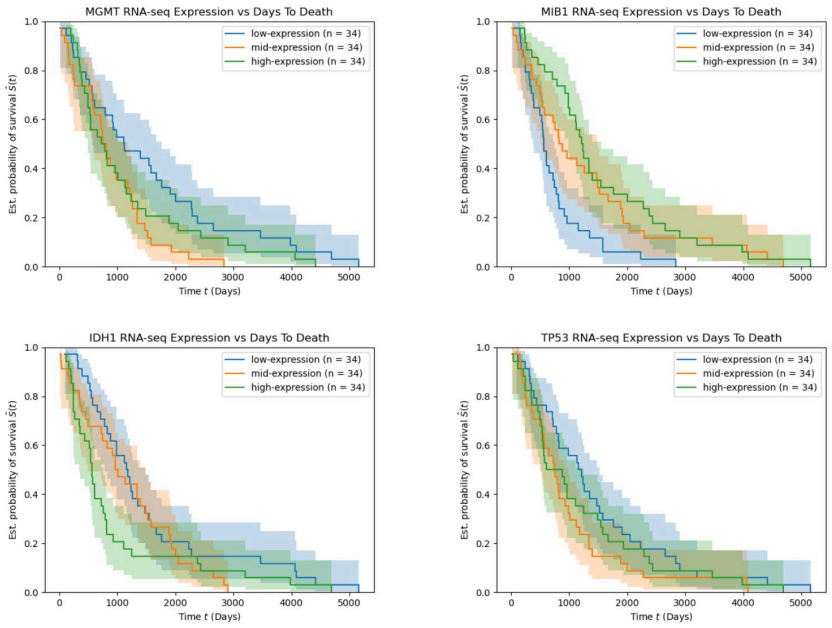

Figure 1: Patient Survival Curves, measured in Days to Death. We stratified the list of recorded patients into 3 categories: low RNA-seq expression (0-33.3 percentile), medium RNA-seq expression (33.4-66.6 percentile), high RNA-seq expression (66.7-100 percentile).

for each of our genes, the three bins (low, medium, and high) RNA-seq expression correspond to different survival curves.

Our exploratory analysis matched the existing literature for six of the seven known cancer-promoting genes (MGMT, IDH1, IDH2, CIC, MKI67, MIB1). In these six genes, the low, mid, and high expression bins showed qualitatively different survival curves. For example, the MIB1 gene showed remarkably different survival curves for low and medium/high expression. We computed the R-squared coefficient of determination for MIB1 to be 0.07, with a p-value of 0.01. This means that with high confidence, MIB1 expression is correlated with cancer survival likelihood, matching the existing literature [3].

The one exception to our analysis was the TP53 gene, which showed very little difference in three survivability curves binned by TP53 expression level. One explanation for this could be that the RNA-seq expression values do not show mutations (changes to the code of the TP53 gene). These mutations, while still affecting patient survivability, do not change the number of copies of the TP53 gene that are made. Therefore, since our RNA-seq data doesn't include mutation information, the survival curves may not appear to be separated in our TP53 plot.

In preliminary model-building, we first applied a Linear Regression model and an un-optimized L2 Regression. It was found that L2 Regression performs much better. Then, we developed a feature reduction heuristics to be applied on both the Linear and L2 Regressions. Lastly, we fine tuned the L2 Regression as it significantly out performed the Linear Regression.

Task 2: Predictive Task and Model Evaluation

Pre-applied Normalization The dataset obtained was already normalized into units of *FPKM*, Fragments Per Kilobase of transcript per Million mapped reads. Essentially, it was normalized by the total number of *reads* in each sample, and normalized by the *length* of each transcript. Under this normalization, features from samples with generally more/fewer reads sequenced are not represented with bias, nor does the length of transcript introduce bias (e.g. a longer transcript will normally be broken into more reads).

Goals and Input Features Two goals are set up: 1) we aim to predict the RNA expression (measured in normalized RNA transcripts sequenced) for a specific gene of interest, and 2) after fitting the model, the fitted parameters (coefficients for the regressor), can also be used to retroactively explain the relationship between literature-supported cancer-associated genes and novel genes. The input to our model will be the normalized RNA expression values (*FPKM*) of the remaining 60659 genes.

In analogy, each patients' RNA-seq data contains the set of ratings they had for each item (RNA). After learning from the many user-item ratings (read-counts), we aim to predict the rating a user would give to a specific item (RNA) given all their previous ratings (other RNAs' read-counts). Furthermore, we would like to explain which set of previous item ratings contributed to the rating we predicted.

The RNA-seq data in *FPKM* was readily available on *The Cancer Genome Atlas* (TCGA). Simply select the cohort of patients and filter by the cancer type would give you the correct subset of data from the database. Our project's dataset information is obtained from the TCGA-LGG (Lower Grade Glioma) Cohort. The survival length information was obtained by downloading the clinical metadata (available on the TCGA-LGG project homepage). For our exploratory analysis, we excluded patients who did not have a recorded date of passing.

Evaluation and Baseline The dataset was split 8:1:1 for training, validation, and testing. The validation set was used to determine the best regularizer and object score setup across all models. The training data is used to determine the number of features within each model and train the final model. Ideally, it would be good to use the validation set to determine the number of features, but due to the noise and smaller sample count, it was found that this particular parameter is better trained with the larger training set.

The mean-predictor was used as baseline predictor. Performance is measured by the MSE of the model on the testing set. Since we are training multiple models, each aimed to predict the quantity of a different RNA, during fine-tuning and evaluation, we normalize each model's MSE by diving the baseline MSE. This gives the fractional error rate relative to the baseline, so performance remains comparable across models.

Task 3: Model and Fine-tuning

Model Selection We have chosen regression models for two main reasons: 1) RNA-seq data are fully filled, so there is no need to infer the missing entries in the dataset, and 2) unlike dimensional reduction models, regression models offer better interpretability.

Training Since each sample contains about 60,000 features, it is necessary to select the correct number of features to prevent overfitting. We optimized three sets of paramters: 1) the regularizer, α, in L2 regressor, 2) the number of features for each model (one RNA count feature we are predicting), and 3) the objective metric used to select the number of features. One major issue we faced and attempted to overcome was the relatively large noise in biological data.

Optimize Regularizer Due to the long runtime, we were not able to perform a full grid search between the regularizer and the objective metric mentioned below. Through early exploratory analysis, $\alpha \approx 1$ works well enough in general. Different α was tried on each of the L2 regression model, and $alpha = 1.15$ seemed to be slightly favored (though any value around 1.0 seemed very similarly qualified).

Reduce Feature Counts We used Recursive Feature Elimination (RFE) to reduce the number of features in each model. RFE is a heuristic to iteratively reduce the number of features included in a regression model. Given a step size s, in each iteration, the model is fitted with the training data, and the s features with the smallest magnitude of coefficient are removed. This is repeated until a desired number of features is reached.

To decrease the runtime, we chose a variable step-size RFE approach, where $s = 100$ is first used to reduce the model to 30,000 features, followed by $s = 20$ until 2,500 features and then $s = 5$ until 500 features. After this, $s = 1$ is used and an objective metric is generated for each number of features

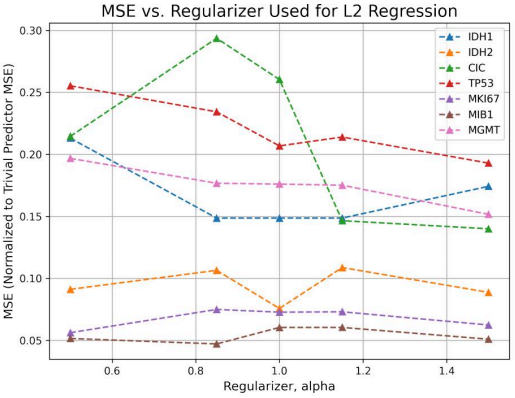

Figure 2: With alpha=1.15, the models have the lowest MSEs

from 500 to 1. The set of features with the best objective metric from the training set is kept to be included in the final model.

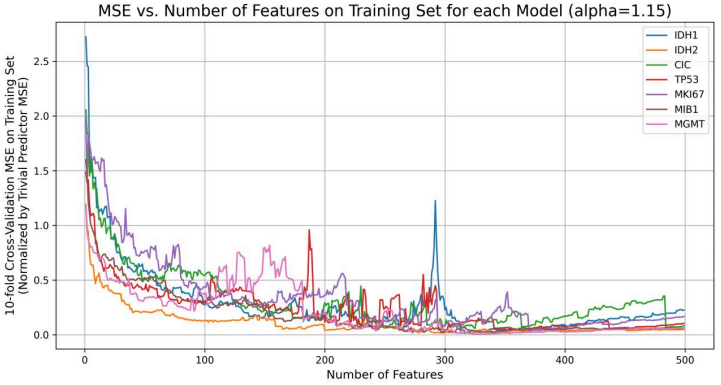

Figure 3: In general, we observed that models have lowest error rate when included slightly more than 300 features

Coping with the Noise The objective metric for each selected set of features from 500 to 1 is computed as a summary statistics of MSEs from a 10-fold cross-validation. Having split into more sets reduces the fluctuation of score due to noise. Initially, mean was chosen as the summary statistics, but it was quickly observed that a more outlier-resilient objective metric was needed. After trying Q1, Q3, Min, Max, Median, and Mean on the validatin set, it was determined that Q3 (third-quantile) performed the best.

Task 4: Relevant Literature

Models Built on TCGA A related work by Padegal et al. [8] used RNA-seq output from a different TCGA dataset, and trained a variety of machine learning architectures (XGBoost [9], Logistic Regression, Feedforward MLP). The authors then fine-tuned each of these models on to predict sample mortality (survival) status to obtain a mean ROC-AUC (area under the receiving curve) score of 0.88 (T. While the authors did not record the self-supervised (RNA-seq expression prediction)

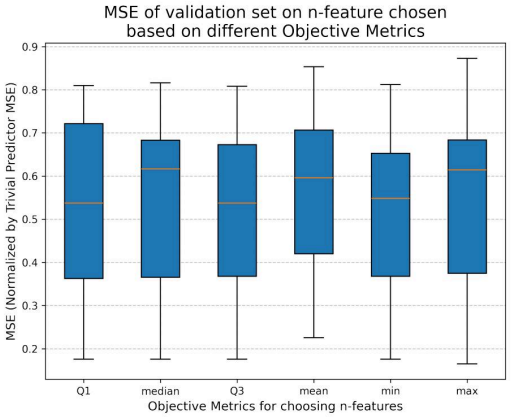

Figure 4: n-features chosen using quartile MSEs performs about 10% better, with Q3 being the best

accuracy, the subsequent fine-tuning analysis is relevant to our model and would be a potential next step.

Regression-Based Models on Biological Datasets Regression models has been widely used in biological data, partially due to high-throughput biological data rarely having missing entries. For example, the work of Morais-Rodrigues et al. trained a logistic regressor on Microarray data to classify breast cancer progression [10]. In the paper of Li et al., it mentioned that feature reduction is necessary as it is often the case for biological data to have more features than the number of samples. This leads to overfitting, and for the case of classifiers, incomplete ranks lead to more than one optimal solutions [11]. Our use of RFE was inspired by this paper, although they applied to an SVM while we did to a L2-regressor. Huge time consumption was a major issue we faced when using RFE to reduce the number of feature, thus we have chosen the naive approach of having larger step-size initially. This paper by Li introduced a combined SVM-RFE approach that greatly speeds up the process. On the other hand, the Morais-Rodrigues paper developed an advanced technique by including "stabilizing term", which expands the feature matrix into a non-singular square matrix, offering a unique solution to the regression-based classfier.

Use of SHAP Score to Enhance Interpretability To determine which of the input features (which genes) are relevant in predicting the RNA-seq expression of our seven known cancer-promoting genes, we computed the SHAP score [12]. As described in Task 5, this score can be used to estimate the role of each input feature in the model prediction. Important features will change the model prediction by large amounts, thereby having a large absolute SHAP score. Using the mean absolute SHAP metric, we rank the top 10 most important genes for each prediction model. This allows us to uncover relevant biomarker genes, and validate our models using existing literature.

Biological Relevance of the Identified Genes We also reviewed the literature to confirm whether the genes predicted to correlate with cancer-promoting genes of interest, as determined by SHAP scores, have biological relevance. Details can be found at the end of Task 5.

Task 5: Results and Conclusion

Our models predicts the RNA-seq expression for each of these 7 known cancer-promoting genes (IDH1, IDH2, CIC, TP53, MKI67, MIB1, MGMT) in *Lower Grade Glioma*. The error rate was evaluated on the testing set with MSE. We see that under the same variable-stepsize RFE setup, the Linear Regression model has significantly less predictive power on RNA expression. Therefore, for the remaining task, we will only focus on using the L2 Regressor. Note, the absolute MSE value does

not matter as the mean count for genes vary greatly. To evaluate the performance of a model, we used the ratio between the model's MSE against the trivial predictor's MSE.

Techniques / MSE for each gene	MGMT	IDH1	IDH2	CIC	TP53	MKI67	MIB1
Trivial Predictor, by Mean	1.1	32.6	440.1	12.1	20.4	7.1	25.4
Linear Regression	0.98	24.1	343.4	6.4	22.8	1.2	5.5
L2 Regression, optimized alpha	0.86	17.4	268.3	6.5	15.0	1.3	5.1

Table 1: MSE of Trivial Predictor, Linear Regressor, and L2-Regressor

More importantly, our second goal is to use our trained model to determine which features affect RNA-seq expression of these 7 genes (which genes are important to predict RNA-seq). To measure this, we employed the SHAP [12] score criterion, which approximates Shapley Values from machine learning interpretability research. The basic idea is that to determine what effect a given feature has on the model are important, we need to see how the model's predictions change if the feature is removed. The exact equation is given below, taken from the Shapely Value page from Wikipedia [13].

$$\begin{aligned}\varphi_i(v) &= \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(n-|S|-1)!}{n!}\left(v(S \cup \{i\}) - v(S)\right) \\ &= \frac{1}{n} \sum_{S \subset N \setminus \{i\}} \binom{n-1}{|S|}^{-1} \left(v(S \cup \{i\}) - v(S)\right)\end{aligned}$$

Figure 5: Exact Shapely Value Equation[13]: Features with high Shapely Values increase the model prediction when included. Features with low Shapely Values decrease the model prediction when included.

One downside is that computing exact Shapely values for each of our hundreds of features is computationally intractable (requiring exponentially many models to be trained). Luckily we used KernelSHAP [12], which is a method of approximating the Shapely values for each of our model features without needing any additional model training.

We applied KernelSHAP to each of our 7 models, and recorded the approximated SHAP scores on the testing set.

Below, here is the KernelSHAP approximation of the exact Shapely Values for our MGMT prediction model.

Figure 6: Example Approximated SHAP Scores for MGMT Prediction. The 10 genes highlighted represent features with high SHAP importance, with their RNA-seq expression output listed. Genes in red represent features that increased the MGMT prediction, and genes in blue decreased the MGMT prediction.

One useful feature of SHAP scores is that we can use the mean SHAP score magnitude to rank how important each feature is in model prediction. In our case, this means for each of our input features (RNA-seq entries), we can take the mean of their absolute SHAP scores. Then, after sorting our list of genes by their mean SHAP magnitude, we can get a list of important genes that are connected to our 7 cancer-promoting genes.

The highest ranked gene our MGMT-predicting model is the SYP gene, which is responsible for making vesicle (transport) proteins in the synapse of neurons [14]. A 2021 paper by Xiao et al. in 'Frontiers in Oncology' found that the SYP gene is highly correlated to the mortality rate of low grade glioma patients [15]. Additionally, this study noted that MGMT is a 'well-known' marker of SYP gene expression!

This is important since it first means that our model is correctly identifying important biomarkers for predicting expressed cancer-promoting genes (MGMT in this example). Second, this means we are able to identify which genes are related to our 7 cancer genes, thus allowing us to create a network of gene expressions.

MGMT Prediction: Top 10 Genes
By SHAP Importance

Gene Name	Mean Shap Magnitudes
SYP	0.30
YWHAH	0.26
SLAIN1	0.22
DNAJB2	0.19
CHGB	0.19
MTURN	0.17
CALM1	0.17
CALM3	0.17
PRKAR1B	0.16
RBM10	0.16

Figure 7: This table displays the top 10 most significant genes in MGMT Model Prediction. We defined significance as the Mean absolute value of SHAP scores for each input feature.

To pick another example: let us take the gene YWHAH, which has a mean SHAP magnitude of 0.26. A recent 2024 paper by Huang et al. identified that YWHAH could play a role in affecting survival of glial (brain) cells [16]. They continue that YWHAH could become a 'pivotal' biomarker in glioma cancer research [16]. We have recorded the ranked SHAP scores for the remaining 6 cancer gene models, with their corresponding ranked important genes.

In conclusion we have determined that our model is able to accurately predict the expression of cancer promoting genes given a filtered set of RNA-seq data. Our models correctly identifies known biomarkers for these cancer promoting genes, as validated by relevant oncology literature. Finally, we can find potentially unknown genes which related to our 7 cancer-promoting genes by using the SHAP importance metric on our trained model. Lastly, as a potential future project, we could use our model to predict downstream tasks such as mortality rate and tumor growth rate, as done by Padegal et al. and Morais-Rodrigues et al., respectively.

Supplementary

CIC Prediction: Top 10 Genes
By SHAP Importance

Gene Name	Mean Shap Magnitudes
UBE2M	1.06
TUBB4A	1.03
PHYHIPL	0.98
MAG	0.78
FBXL16	0.77
RPS19	0.70
FAU	0.61
FCHSD2	0.57
RPS13	0.56
NCDN	0.55

IDH1 Prediction: Top 10 Genes
By SHAP Importance

Gene Name	Mean Shap Magnitudes
RPL32	2.21
NDUFB10	1.86
RNASE1	1.63
TNR	1.59
SULT4A1	1.53
PTTG1IP	1.40
PLAAT4	1.32
SRSF5	1.25
SNAP25	1.22
ADGRG1	1.20

IDH2 Prediction: Top 10 Genes
By SHAP Importance

Gene Name	Mean Shap Magnitudes
ZNF710-AS1	4.57
YWHAG	4.38
SULT4A1	2.75
RPLP2	1.96
ALDH4A1	1.88
LAMTOR4	1.67
RTN3	1.65
GRN	1.64
NCAM1	1.56
EEF2	1.53

MIB1 Prediction: Top 10 Genes
By SHAP Importance

Gene Name	Mean Shap Magnitudes
RPRD1A	0.46
SYT1	0.43
FCER1G	0.40
MOG	0.39
SPIRE1	0.34
CPE	0.31
ANGPTL2	0.30
RPLP0	0.30
PTPRZ1	0.30
RPL7	0.30

MKI67 Prediction: Top 10 Genes
By SHAP Importance

Gene Name	Mean Shap Magnitudes
TPX2	0.49
CENPF	0.32
RPS25	0.30
RPS8	0.28
TOP2A	0.27
GPM6A	0.24
PHPT1	0.18
CPLX2	0.17
SCRN1	0.16
CDCA5	0.16

TP53 Prediction: Top 10 Genes
By SHAP Importance

Gene Name	Mean Shap Magnitudes
RPL35	1.26
RPL4	1.06
TOMM7	1.00
PSMB1	0.89
CNP	0.88
ENC1	0.85
RPL35A	0.83
ARPP19	0.70
RPL11	0.67
ELOB	0.64

Figure 8: This table displays the top 10 most significant genes for each Model Prediction task. We defined significance as the Mean absolute value of SHAP scores for each input feature.

References

[1] John N. Weinstein, Eric A. Collisson, Gordon B. Mills, Kenna R. Mills Shaw, Brad A. Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M. Stuart. The cancer genome atlas pan-cancer analysis project. 45(10):1113–1120. Publisher: Nature Publishing Group.

[2] Michael Weller, Roger Stupp, Guido Reifenberger, Alba A. Brandes, Martin J. van den Bent, Wolfgang Wick, and Monika E. Hegi. MGMT promoter methylation in malignant gliomas: ready for personalized medicine? 6(1):39–51.

[3] David Gorodezki, Julian Zipfel, Nägele Thomas, Martin Ebinger, Martin U Schuhmann, and Jens Schittenhelm. LGG-01. PROGNOSTIC UTILITY AND CHARACTERISTICS OF MIB-1 LABELING INDEX AS a PROLIFERATIVE ACTIVITY MARKER IN CHILDHOOD LOW-GRADE GLIOMA: A RETROSPECTIVE OBSERVATIONAL STUDY. 26:0.

[4] Sourat Darabi, Joanne Xiu, Timothy Samec, Santosh Kesari, Jose Carrillo, Sonikpreet Aulakh, Kyle M. Walsh, Soma Sengupta, Ashley Sumrall, David Spetzler, Michael Glantz, and Michael J. Demeure. Capicua (CIC) mutations in gliomas in association with MAPK activation for exposing a potential therapeutic target. 40(7):197.

[5] Adam Cohen, Sheri Holmen, and Howard Colman. IDH1 and IDH2 mutations in gliomas. 13(5):345.

[6] Humaira Noor, Nancy E. Briggs, Kerrie L. McDonald, Jeff Holst, and Orazio Vittorio. TP53 mutation is a prognostic factor in lower grade glioma and may influence chemotherapy efficacy. 13(21):5362.

[7] Shi-yi Wu, Pan Liao, Lu-yu Yan, Qian-yi Zhao, Zhao-yu Xie, Jie Dong, and Hong-tao Sun. Correlation of MKI67 with prognosis, immune infiltration, and t cell exhaustion in hepatocellular carcinoma. 21:416.

[8] Girivinay Padegal, Murali Krishna Rao, Om Amitesh Boggaram Ravishankar, Sathwik Acharya, Prashanth Athri, and Gowri Srinivasa. Analysis of RNA-seq data using self-supervised learning for vital status prediction of colorectal cancer patients. 24(1):241.

[9] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

[10] Francielly Morais-Rodrigues, Rita Silverio-Machado, Rodrigo Bentes Kato, Diego Lucas Neres Rodrigues, Juan Valdez-Baez, Vagner Fonseca, Emmanuel James San, Lucas Gabriel Rodrigues Gomes, Roselane Gonçalves dos Santos, Marcus Vinicius Canário Viana, Joyce da Cruz Ferraz Dutra, Mariana Teixeira Dornelles Parise, Doglas Parise, Frederico F. Campos, Sandro J. de Souza, José Miguel Ortega, Debmalya Barh, Preetam Ghosh, Vasco A.C. Azevedo, and Marcos A. dos Santos. Analysis of the microarray gene expression for breast cancer progression after the application modified logistic regression. *Gene*, 726:144168, 2020.

[11] Zifa Li, Weibo Xie, and Tao Liu. Efficient feature selection and classification for microarray data. *PLOS ONE*, 13(8):1–21, 08 2018.

[12] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions.

[13] Shapley value. Page Version ID: 1256327414.

[14] PubChem. SYP - synaptophysin (human).

[15] Zheng Xiao, Shun Yao, Zong-ming Wang, Di-min Zhu, Ya-nan Bie, Shi-zhong Zhang, and Wen-li Chen. Multiparametric MRI features predict the SYP gene expression in low-grade glioma patients: A machine learning-based radiomics analysis. 11:663451.

[16] Ruiting Huang, Ying Kong, Zhiqing Luo, and Quhuan Li. LncRNA NDUFA6-DT: A comprehensive analysis of a potential LncRNA biomarker and its regulatory mechanisms in gliomas. 15(4):483.