DATA SCIENCE

Exploring Liver Cancer Classification and Survival Prediction with Multi-omics Analysis

Blessy Rajan*†, Gabriela Djuhadi† and Shubhangi Kaushik†

*Correspondence:
bler98@zedat.fu-berlin.de
Freie Universitat, Berlin, Germany
Full list of author information is available at the end of the article
†Equal contributor

Abstract

The survival rate for the 43% of people getting diagnosed in the early stage of Liver cancer is 36% which drops with the progression of the disease. The project tried to predict biomarkers relevant to the disease and their contribution to the survival rate. The goal of this project was to perform feature reduction using differential expression analysis and machine learning analysis for multi-omics data and conduct survival analysis. In the end, the important features were obtained through the two different feature reduction methods, and the biological relevance of those to liver cancer was found. Scores and ROC curves were developed for these methods and compared to one another. Survival analysis was conducted on the top features obtained by the best performing model. Features were reduced by differential expression analysis through DESeq, Limma, and Logistic regression. Support vector machine(SVM), Gradient boost(GB) model were built for these features classifying if the patient was alive or dead. Survival analysis was done using Cox regression, Keplen-Meier Curve and Weibull analysis. The project focuses on the prediction of important biomarkers and their contribution to the survival rate. This could support the diagnosis of the patient by also providing the data on the biomarkers contributing the most to the disease and survival rate. This approach could help clinics to provide personalized treatment to the patients.

Keywords: Liver Cancer; Differential Expression Analysis; Machine Learning; Data Science; Multiomics Analysis; Survival; mRNA; miRNA; Methylation; CpG

Background

Liver cancer, in particular the Hepatocellular Carcinoma (HCC), the most prevalent type of said cancer, is a malignant tumor that begins in the liver. It is particularly aggressive, ranking third as a cause of cancer death in males [1]. As a heterogeneous condition with various confounding and risk factors[2], many patients are diagnosed after showing advanced disease. Staging and prognosis remain a challenge, and the limited treatment strategies create a pressing need for a reliable survival prediction tool [3].

A previous study by [2]. considers a variety of identified molecular subtypes of HCC. They gathered the data from the TCGA multi-omics cohort: the miRNA expression, mRNA expression, methylation data, and the clinical data, to predict the survivability of the patient across all subtypes and stages. They constructed a complex unsupervised neural-network algorithm, called Autoencoder, to reduce the dimensionality of the combined data through clustering. The three different omics data were first merged, ran through autoencoder, and then used to train a Support

Vector Machine (SVM) classifier to predict whether the patient will survive or not.

Due to the nature of the clustering method used, feature importance analysis on the predictions could not be performed. This omits the possibility of finding possible biomarkers for a more accurate prognosis and earlier diagnosis, which could have incited follow up experiments or aid in that line of research. Additionally, it was considered that early-clustering approach, i.e. where the data were merged before clustering, may disproportionately weigh one type of omic data over the others simply from differences in variance, which may affect the classifier's decisions later on.

In this line, our project examined two different feature reduction methods, coupled with two different classifiers, and finally performed feature importance and survival analysis as an attempt to find potential indicators of liver cancer survivability. Owing to the previous reasoning, we used late-clustering for the feature selection with machine learning as well as feature selection with differential expression analysis. The two classifiers chosen were Support Vector Machine (SVM) and Gradient Boost (GB) estimator, out of which we were able to extract the important features to find further biological context and perform the survival analysis.

Goal

One of the goals of the project was to analyze the influence of the different preprocessing methods, on machine-learning-based predictions. As mentioned in the background, the early clustering method for reducing the dimensionality of the data i.e. the feature reduction process used in the previous study, could be more partial towards omics data containing more data compared to the other omics data to which it merged. Hence one of the goals of the study was to explore other ways to reduce features of the multi-omics data while preventing biases towards the omics dataset type with more data. The methods employed here for the feature reduction are a machine-learning-based approach using logistic regression and an R-based differential expression analysis. Later on, ROC curves were obtained and scores were noted for the data from both methods using Support Vector Machine(SVM) and Gradient boost. On top of the top 100 features obtained from the best working model out of the combination of the 2 preprocessing steps and the 2 machine learning algorithms, survival analysis was conducted and the important biomarkers pertaining to liver cancer were analyzed for their biological relevance.

Data and Preprocessing

The project utilized liver cancer data from TCGA, specifically retrieved using the LIHC code. While the mRNA and miRNA data were successfully downloaded without any complications, challenges arose when attempting to download and open the methylation data. Consequently, three alternative approaches were pursued:

1. Continuing the analysis by solely considering miRNA and mRNA data.
2. Exploring alternative methods to download the methylation data.
3. Find a new research paper with usable multi-omics analysis.

After failing approach 1 and approach 2, the team started looking for a new dataset to work with. In the process, a research paper[4] was discovered that matched the project requirements. Furthermore, during the examination of the provided data, we eventually realized that all the pertinent information relevant to the initial research topic[2] was present in the paper. The research paper talked about using preprocessed data, raising concerns about its impact on our results. However, later it was discovered they had excluded some of the redundant data columns from the mRNA and miRNA datasets In mRNA data, the column considered for the analysis was median length normalized data, and for miRNA data, the column considered for the analysis was reads-per-million-miRNA-mapped. For methylation data, the data consisted of beta values which were later converted to M values. Overall, we were dealing with original data only which consisted all the relevant information for the three datasets. On the contrary, survival data retained was too less columns compared to the clinical data from TCGA, making the latter more appropriate for the overall analysis.

The types of data used for preprocessing were miRNA, mRNA, Methylation, and Survival. The number of samples for survival data was 438, while for miRNA, mRNA, and methylation data, it was 425, 424, and 430, respectively. As one of the focus points of the project was to conduct survival analysis, it was considered to take samples from other omics data which were common in the survival data. Hence, The data found to be common between the methylation and survival data were 423 samples, the miRNA and survival data had 418 samples in common, and the mRNA and survival data had 417 samples in common. However, altogether the data which was common between all 4 dataset types were 404 samples. While filtering out the common data, one of the considered criteria was the presence of death and survival days in the survival data. The samples which did not contain data in these columns were excluded from the analysis.

During data exploration of clinical data, it was observed that the survival days varied up to approximately 3200 days, but the majority of patients survived only up to 1000 days (Figure 10). The data exhibited an imbalance between the deceased and surviving patients, with a larger proportion of data belonging to the surviving class. Particularly, the male class showed a bias towards more surviving data. For tumor stage data, stages 1 and 2 had an overrepresentation of surviving patients (Figure 11) and understandably so considering patients tend to have higher chances of surviving the initial years with cancer. This issue was addressed through under-sampling. Additionally, subclasses of stages 3 and 4 were merged into single subclasses. Surprisingly, after these adjustments, both the issue of more surviving data and the bias in the male class were resolved, resulting in a balanced dataset ready for analysis.

Differential Expression Analysis with R

The purpose of differential expression analysis is to obtain the genes or transcripts that have significantly varying expression levels across the samples. Hence, this analysis helped in obtaining the genes or transcripts contributing the most to the

disease. While working with mRNA and miRNA the data needed for this analysis is the read counts data, describing the level of expression of the probes. For the methylation data, the data required for this analysis is the m-values or methylation values. The methylation data obtained consisted of beta values, which could not be used for analyzing the differential expression, as beta values provided more intuitive interpretation whereas the m values are preferred more for the statistical analysis[5]. The libraries used for this analysis were R-based libraries, and the ones finally used were DESeq and Limma; other libraries were also explored nonetheless because of not being able to obtain good results, they were discarded. For the DESeq analysis, the data should be integer values, and the negative binomial is used to obtain the differential expressed data, whereas, Limma requires linear modeling and empirical Bayes methods for differential expression analysis. While conducting the analysis of miRNA data in DESeq the differential expression data was obtained at 0.1 as p-value and at 1 as Log 2 fold change value. For mRNA analysis in DESeq, the p-value was set at 0.05, and Log 2 fold change at 1. While conducting the analysis for methylation data the beta values were converted into m values using the formula:

$$Mvalues = log_2(\frac{BetaValues}{1 - BetaValues}) \tag{1}$$

Later those m values were used for the analysis at 0.1 as the p-value and log 2-fold change value at 1. During the methylation data analysis, the data was partitioned into 3 parts, and all of them were analyzed separately and the distinct results were appended together. At the end of the analysis, 181 mRNA probes were found to be differentially expressed, from miRNA 4 probes were found to be differentially expressed, and from methylation data, 24 probes were found to be differentially expressed. Eventually, the total features got reduced, and these data were merged and used for the machine learning analysis to obtain important biomarkers.

Logistic Regression

Feature selection with Machine Learning was done using Logistic Regression as opposed to Linear Regression based methods considering that the classification that follows was binary. This was implemented using sklearn's LogisticRegression and SelectFromModel methods, which takes the β values of the logistic function of the following form after being fitted to the data.

$$p(X) = \frac{e^{\beta_0} + e^{\beta_1}X_1 + ... + e^{\beta_n}X_n}{1 + e^{\beta_0} + e^{\beta_1}X_1 + ... + e^{\beta_n}X_n} \tag{2}$$

Where n is the number of features in the data, β being the assigned weights, and $p(X)$ is the probability that the featured datapoint is of class 1. The loss function for this model, given that the regularization function is l1, follows the form below.

$$LossFunction = \frac{1}{n}\sum_{i=1}^{n}(\hat{Y_i} - Y_i)^2 + \lambda\sum_{i=1}^{N}|\theta_i| \tag{3}$$

Where $\hat{Y}$ is the predicted value of the regression, a part of the squared error function, while θ is the weight (in this case the β values) for each feature and λ is the

weight constant, a part of the regularization function [6]. Since l1 regularization shrinks and may reduce the weights to 0, SelectFromModel with this regularization returns all the features where the β value is at least 10^{-5} by default. This logistic regression model is tuned by the parameter C, which is the inverse of regularization strength, or how much error from the actual data point can be tolerated.

For each omic data, a LogisticRegression model was constructed, the data were processed by the model, concatenated, split into train and test, and then fed into a Gradient Boost estimator to find the combination of C parameters that yields the best test AUC. This was exhaustively iterated with different C values ranging from 0 to 1 incremented by 0.1 for each omic data. The best combination was found to be 0.1 for mRNA, 0.1 for methylation, and 0.4 for miRNA.

This search was only for the C value combination; the Gradient Boost estimator used the best parameters obtained from hyperparameter tuning with the data processed with R at the time. This search was greatly time consuming, and thus after the best Gradient Boost parameter was updated and modifications to the data distribution was done, there was not enough time to run this again. The data processed with this method only underwent random removal for stage 1 and 2 as well as dropout removal based on the clinical data. Further balancing was done with Synthetic Minority Oversampling Technique (SMOTE) and RandomUnderSampler methods from sklearn.

0.1 Normalization and Scaling

One of the classifiers used in this project is the Support Vector Machine (SVM) with a linear kernel. This classifier would only work well if the data can be separated by a linear hyperplane, requiring examination of the data distribution. In doing this, the quantile-quantile (QQ) plot against normal distribution was generated for the three datasets.

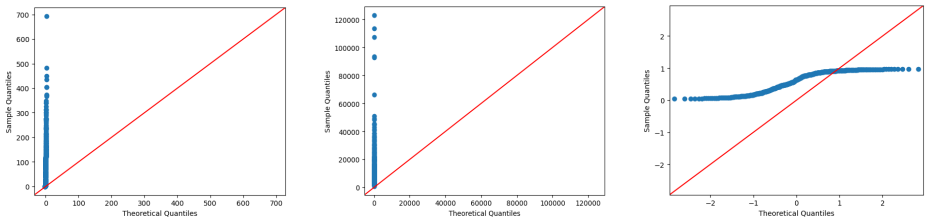

Figure 1: QQ-plots of the raw mRNA (left), miRNA (middle), and methylation (right) data.

The figure above suggests that the values of mRNA and miRNA followed exponential distribution since they are spread far apart, and the methylation values follows a bimodal distribution. Because of this, log transform applied to the mRNA and miRNA data, while Quantile Transform method from sklearn was used for the methylation data. Quantile Transform is a non-linear transform that is applied to each feature independently. It tends to spread out the most frequent values and

reduce the effects of outliers. All the transformed data were then normalized using StandardScaler method from sklearn, which simply subtracts the values by the mean and divides it by the standard deviation independently for each feature.

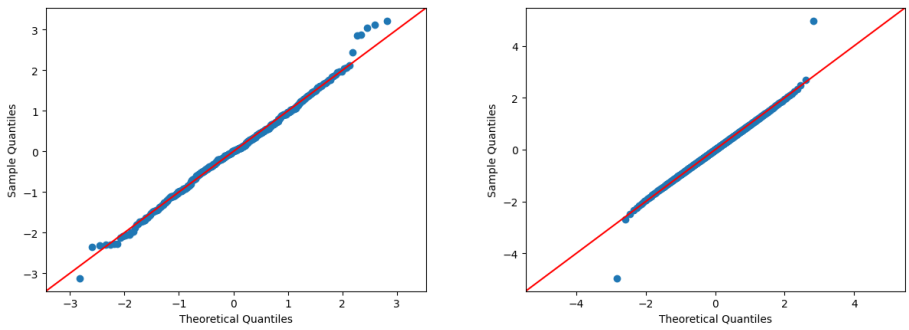

Figure 2: QQ-plots of the processed mRNA (left) and methylation (right) data. The plot for processed miRNA data closely follows mRNA.

This process was done before feature selection with Logistic Regression, but after feature selection by differential expression analysis. It was not done until troubleshooting had to be done to improve SVM performance, and was found to be necessary towards the end of the project.

Methods

After feature reduction and transformation, the individual datasets were merged into a single dataset, incorporating essential features from miRNA, mRNA, methylation, and clinical data, which was then standardized again to maintain consistent scaling across all numerical features. The preprocessed and standardized dataset was used to train data science models, SVM and Gradient Boosting models, with optimal hyperparameters selected based on AUC scores. Model performance was thoroughly analyzed using various evaluation metrics like accuracy, precision, recall, and F1-score. The best-performing models were selected for further analysis. Top 100 features from these models were chosen for survival analysis to explore their association with cancer subtypes and patient survival. Survival analysis employed Cox regression and Kaplan-Meier curves to assess the impact of these features on patient survival, and Weibull analysis was conducted to estimate the number of days left for a person to live, taking into account the influence of significant features on patient survival.

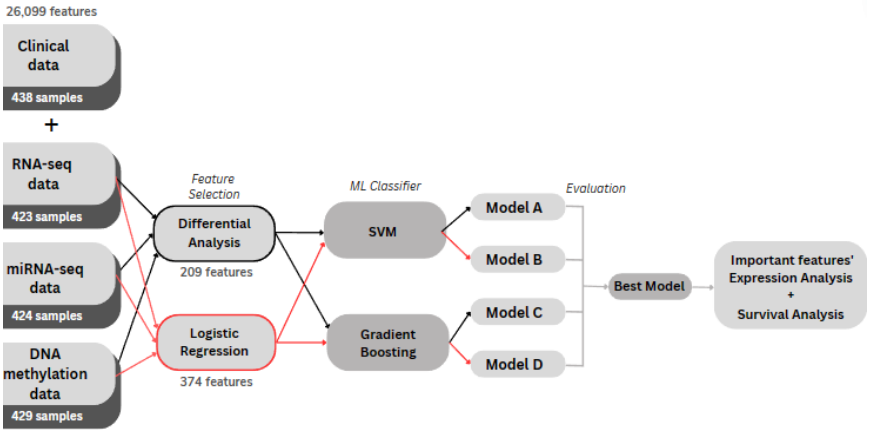

Figure 3: Overall structure of the analysis.

0.2 Machine Learning Models

0.2.1 Support Vector Machine (SVM)

SVM classifier is used in this project, not simply because the paper by Chaudary et. al[2]. used it, but also because it fits the use of this project. SVM does well in higher dimensional space, especially when the number of features is larger than the samples, which is true for our data even after feature reduction [7]. It works by finding the separating hyperplane between the two classes, but with a soft margin to avoid overfitting. It is typically used for binary classification and often does poorly for multi-class problems because the boundaries may overlap and the distribution may not be separable even with non-linear kernel or dimension addition. Inline with this soft margin is the term support vector, which are the data points that lie within the margin, either on the side of its own class or on the other side of the hyperplane. The hyperplane placement is decided by this support vector and the "softness" of the margin, which relates to the parameter C as explained in the following equations. Note that in these equations the variables n and p are switched from the equations in lecture.

$$
\begin{aligned}
f(x_{1,\dots n}) = \beta + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} \ , \quad & \sum_{j}^{n} \beta_j^2 = 1 \\
y_i(f) \geq M(1-\epsilon_i) & \\
\epsilon \geq 0 \ , \quad \sum_{i=1}^{p} \epsilon_i \leq C &
\end{aligned}
\tag{4}
$$

where n is the number of features, ϵ is the error score of that observation, and $f(x_{1,\dots n})$ being the function of the hyperplane. β is the parameter of the said function with a constraint shown above. y_i takes that the result of that function for any data point, and returns 1 for if the $f(\vec{x}) > 0$, and -1 if $f(\vec{x}) < 0$. The amount of ϵ allowed would then be constrained by C, with p being the number of training observations and one observation indicated with i.

Higher numbers of C allows more violations, and therefore can include more observations or support vectors, and the opposite is true with lower C. This implies that larger amounts of C would allow more bias and the smaller would allow more variance into the model. The hyperplane placement would then become a problem to maximize M.

In this project, the linear kernel was chosen for the SVM implementation from sklearn, since it performs well and the feature importance calculation is already implemented. The bias-variance trade-off was determined by trying a range of values of C from 0.1 to 1 by 0.1 increments and finding the best test AUC, which was found to be 0.1.

0.2.2 Gradient Boosting

Gradient Boosting[8] trains weaker algorithms like Decision Trees[9] sequentially, such that every later model corrects the mistakes made by the former model. To achieve this, a loss function is used to measure the error between the model's predictions and the actual target values. The goal is to minimize this loss function during the training process, which guides the optimization of weak learners to improve the model iteratively. The optimization in Gradient Boosting is performed in the function space, where the function estimate $\hat{f}_i(x)$ is parameterized in an additive functional form:

$$\hat{f}(x) = \hat{f}_M(x) = \sum_{i=0}^{M} \hat{f}_i(x) \tag{5}$$

where M is the number of iterations, $\hat{f}_i(0)$ is the initial guess and $\{f_i\}_{i=1}^{M}$ are the function increments or "boosts."

Gradient Boosting applies regularization techniques that help mitigate the chances of overfitting by Decison Trees and overall improve the model's performance. For the models in the project, the default loss function "deviance" was used, which corresponds to the Log Loss or Cross-Entropy Loss. The Log Loss function is given by:

$$\text{Log Loss} = -\sum_{i=1}^{N} \left(y_i \log(p_i) + (1 - y_i) \log(1 - p_i)\right) \tag{6}$$

where N is the number of data points in the dataset, y_i is the true binary label (0 or 1) for the ith data point, and p_i is the predicted probability of the positive class (class 1) for the ith data point.

0.3 Survival Analysis

0.3.1 Cox regression

Cox regression[10], also known as the proportional hazards model, is a statistical method used to analyze time-to-event data in medical research and epidemiology. The primary goal is to understand the relationship between covariates and the likelihood of an event occurring over time. The model estimates the hazard rate

function, representing how the risk of an event changes over the observation period based on the values of covariates. It accounts for censored data where the event time is unknown for some individuals, unlike regression. The Cox regression model is built upon the hazard function. In mathematical terms, the Cox model is represented as follows:

$$H(t) = H_0(t) \times \exp(b_1x_1 + b_2x_2 + \ldots + b_kx_k). \tag{7}$$

Here, $x_1, x_2, \ldots, x_k$ denote the predictor variables, and $H_0(t)$ is the baseline hazard at time t, which signifies the hazard when all predictors are set to zero. By directly obtaining the regression coefficients $b_1, b_2, \ldots, b_k$ from the software, we can calculate the Hazard Ratio (HR) of a specific risk factor or predictor in the model.

This method is widely used to investigate factors that influence the occurrence of specific events, such as the onset of a disease or mortality rates, providing valuable insights for researchers in epidemiological studies.

0.3.2 Kaplan-Meier curves

Kaplan-Meier analysis, a statistical method used in medical clinical trials and other fields, examines changes over time concerning a specific event, such as death or disease progression. It addresses incomplete observations caused by censoring, where some participants are lost to follow-up before experiencing the event of interest. This analysis estimates survival probabilities over time, generating survival curves that visually represent the proportion of participants yet to experience the event. By comparing these curves across treatment arms or groups, researchers evaluate treatment effectiveness and derive meaningful conclusions. The method relies on assumptions of non-informative censoring and independence among participants. To quantify and compare the relative risks of event occurrence between groups, hazard ratios and the log-rank test are utilized. In conclusion, Kaplan-Meier analysis is a valuable tool for understanding time-to-event data and facilitating evidence-based practices.

0.3.3 Weibull analysis

The Weibull model is employed as a fully parametric approach in survival analysis, frequently used in conjunction with Cox's regression. Both models assume proportional hazards and provide hazard ratios for treatment comparison. Going beyond proportionality, the Weibull model also functions as an accelerated failure-time model (AFT), allowing the estimation of event time ratios alongside hazard ratios. Thus, it was used in the project to determine the expected number of days patients may survive. It provides valuable insights even when there are modest deviations from the ideal distribution. Its predictive ability and description of treatment effects in terms of hazard and relative survival time make it an invaluable tool for analyzing clinical trial data. In conclusion, the Weibull model is a powerful and flexible tool for survival analysis, facilitating enhanced patient care and treatment optimization.

Results

Differential Expression Analysis Classification Results

The differentially expressed data obtained using R was fed into SVM and GB to assess which model performed better with the data to predict the important biomarkers. While conducting the hyperparameter tunning to find out the number of estimators at which the GB has the optimum performance, The AUC Score ranged between 60-80%. The number of estimators at 150-170 the GB performed the best, the max_features parameter was set at 3 and random state at 10. The scores for the training and testing set are noted down in the table 1

	GB		SVM	
Indexes	**Training**	**Testing**	**Training**	**Testing**
Accuracy	0.953	0.636	1.000	0.657
Precision	0.956	0.653	1.000	0.660
F1-score	0.952	0.640	1.000	0.673
Specificity	0.948	0.627	1.000	0.686
Sensitivity	0.956	0.645	1.000	0.625

Table 1: GB and SVM performance scores after Differentially expressed analysis feature selection

After obtaining the scores from both models ROC curve was generated to visualize the performance of the models. In both models developed the training set scored very well, however, the scores of the testing set were really low. This suggested the overfitting of the model, i.e. although the model is able to capture the underlying pattern of the training set, nonetheless, it is not able to fit itself to the underlying pattern of the testing set.

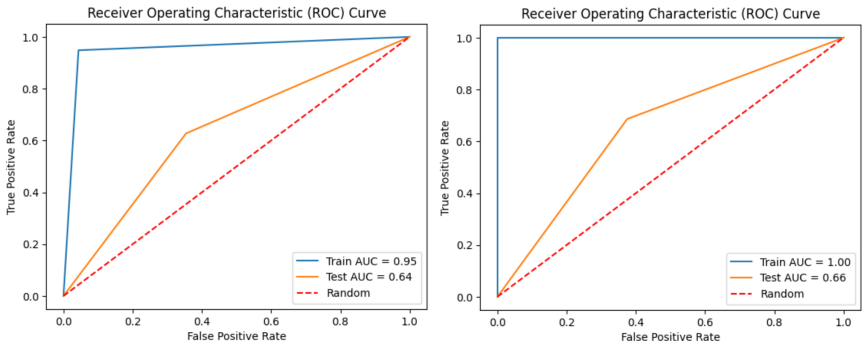

Figure 4: The Receiver Operating Characteristic and Area Under the Curve for GB estimator(left) and SVM (right)

The prediction analysis was conducted multiple times and the overall, gradient boost gave better results than SVM. Hence feature importance was performed and important biomarkers were predicted using the gradient boost model.

The topmost important feature obtained in Figure 4 is NCRNA00162, while going through the research papers to find the biological context of this mRNA, it was found to be directly connected to the progression of liver cancer[11]. FRMD1 is related to FERM Domain Containing protein 1, which gets misregulated and promotes HCC tumorigenesis and metastasis[12]. AHRR, one of the genes for cgisland

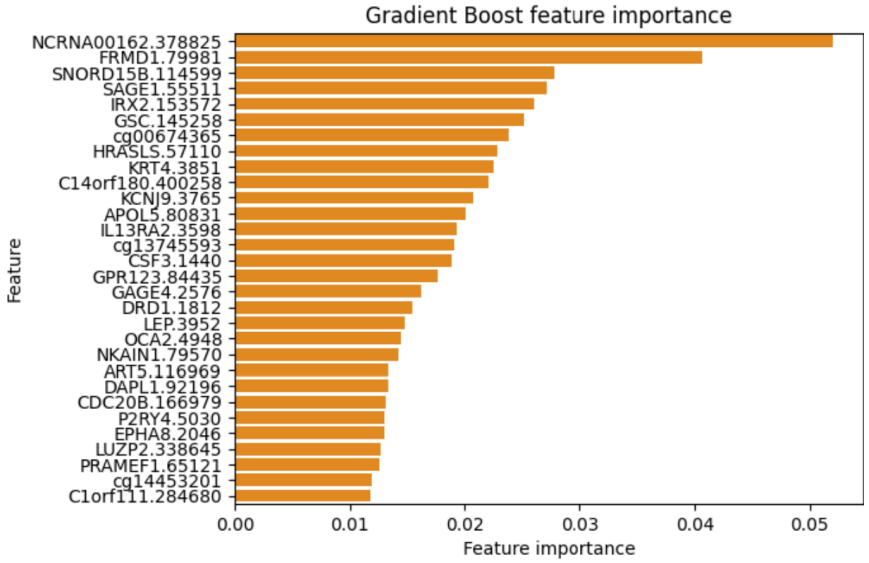

Figure 5: The list of top 30 important biomarkers predicted using gradient boost method

in methylation data was found to be highly correlated to the survival rate of liver cancer patients. For the miRNA data SNORD, the gene for mir1280 was found to be a highly unstable gene in liver cancer[13]. This analysis showed that the important features obtained from this analysis showed high relevance to liver cancer.

Logistic Regression

The feature selection using logistic regression yielded 73 features for methylation data, 219 for miRNA data, and 82 for mRNA data from starting numbers of 5,000, 20,531, and 1,046 respectively. From this, using a GB estimator of 1500 estimators and maximum features of 8, as well as SVM classifier of C = 0.1, the values in the table below were obtained.

	GB		SVM	
Indexes	**Training**	**Testing**	**Training**	**Testing**
Accuracy	1.000	0.826	1.000	0.991
Precision	1.000	0.780	1.000	1.000
F1-score	1.000	0.796	1.000	0.989
Specificity	1.000	1.000	1.000	1.000
Sensitivity	1.000	0.812	1.000	0.979

Table 2: GB and SVM performance scores after Logistic Regression feature selection

It can be seen that the GB estimator is highly overfitted, reaching great performance in the training data, but terrible in the test. It should be noted that the specificity stays the same at 1.000 although everything else suffers, which shows that the classifier recognized all the data points in the negative class. In contrast, SVM did very well in both training and testing data, where the performance was only slightly worse in the testing than the training set.

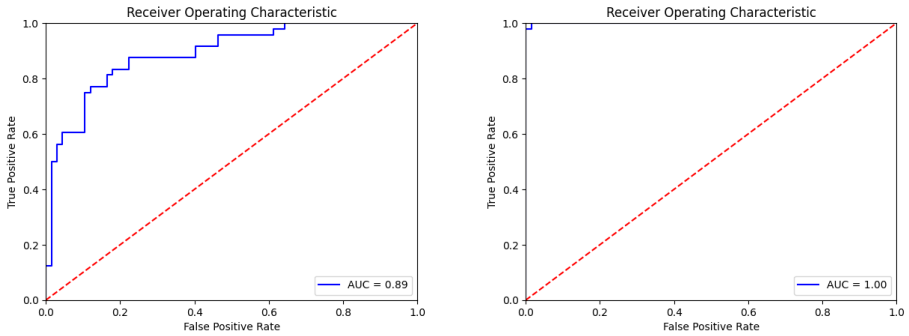

Figure 6: Receiver Operating Characteristic and the Area Under the Curve for GB estimator (left) and SVM (right).

It could be seen in the figure above that the performance of GB estimator is relatively not bad, although the accuracy is only 0.82, the AUC reaches almost 0.9. Note that ROC uses the probabilities of the positive class, which GB estimator does not classify well.

After the feature importance was extracted, the top 5 features of each omic data are listed in the table below.

miRNA

feature	importance
ADAMTS5.11096	0.028655
TAF3.83860	0.026099
PM20D1.148811	0.025672
HPS3.84343	0.025624
PCDHGB2.56103	0.0255

meth

feature	importance
cg23696472	0.023438
cg05213296	0.021056
cg07062847	0.021039
cg25459558	0.013565
cg00084338	0.013453

RNA

feature	importance
hsa.mir.212	0.028355
hsa.mir.215	0.027042
hsa.mir.658	0.015113
hsa.mir.149	0.010693
hsa.mir.548e	0.008488

miRNA

feature	importance
X..100133144	0.009777
ABP1.26	0.007423
ACOXL.55289	0.021113
ACSM4.341392	0.040499
ADAMTS5.11096	0.030658

meth

feature	importance
cg00084338	0.009414
cg00114913	0.003468
cg00187686	0.016258
cg00674365	0.015707
cg00958884	0.023815

RNA

feature	importance
hsa.mir.106a	0.012572
hsa.mir.1185.1	0.018532
hsa.mir.1238	0.018126
hsa.mir.126	0.004778
hsa.mir.1260	0.01573

Figure 7: Top 5 most important features of each omic data from GB estimator (top) and SVM(bottom).

The numbers indicate the importance of that feature for its respective classifier, and this cannot be cross-compared. The omics data were not equally distributed in the sorted feature importance, as indicated in Figure 1 in the appendix, where the top 18 features were dominated by miRNA, which also has the most features com-

pared to the other two omics data. Between the two classifiers, whether in this top 5 table for each omics data or for the top 18 overall, only two features are common in both classifiers: ADAMTS5.11096 and cg00084338. This could be a marker for HCC survivability. After research through Ensembl database, cg00084338 is located at Chromosome 6: 170,286,832-170,286,881 [14]. It is located close to a promoter and transcribed regions, but known variants are only associated with intronic modification of the transcript ENST00000630384.2. Being an epigenetic marker, it is possible that it does not affect gene expressions directly in that region, yet possibly in another region. Unfortunately, there was no interaction data available for this genomic region nor this CpG island.

ADAMTS5 with gene ID: 11096, known officially as ADAM metallopeptidase with thrombospondin type 1 motif 5, is often associated with cartilage and osteoarthritis, and is not as highly expressed in the liver as it is in other tissues such as in the uterus, brain, thyroid, and spleen. Nonetheless, there was an in-silico study conducted by Zhu et. al [15]. that reveals the positive correlation between ADAMTS and HCC development, and significantly so with a worse survival rate with p ¡ 0.05. This study examined specifically HCC, pursuing the possibility of using this marker as an indicator for prognosis by performing Kaplan-Meyer analysis, univariate, and multivariate COX analysis. As opposed to the multi-omics approach of this project and that by Chaudhary et. al., this paper only examines mRNA data, combining several datasets from GSE and TCGA. A subset of the data from GSE was used by Chaudhary et. al. as a confirmation cohort, but the data obtained from TCGA contained more samples that the data used in this project could be entirely included there. Zhu et. al. also conducted GO enrichment analysis, gene set enrichment analysis (GSEA), and weighted gene co-expression network analysis (WGCNA), and found that ADAMTS5 could be involved in converting a cancer from metabolic-dominant to extracellular-dominant. This status conversion is likely to cause higher mortality risk in patients(Figure 13).
Another study found that ADAMTS5 gene is significant for colorectal cancer (CRC) prognosis [16]. They have found that hypermethylation of that gene is strongly associated with longer recurrence-free survival (RFS). This is interesting considering CRC is somewhat related to liver cancer, where 25%-50% of patients with CRC develop colorectal liver metastases (CRLM)[17].

The top 100 most significant features were extracted from the Gradient Boost analysis done on the data reduced by Bioconductor packages, along with all of the clinical data. Having successfully found the most significant features, the next pivotal step in the analytical pipeline involved the application of survival analysis techniques. Specifically, Cox regression, Kaplan-Meier curve analysis and Weibull analysis were used on the refined set of selected features. These methods aided in gaining deeper insights into the underlying survival patterns and time-to-event outcomes present in the data.

After applying the Cox model to the top features identified by the gradient boost model on data reduced by differential analysis, we obtained favorable results with

a concordance index of 0.84, indicating good predictive accuracy, and a well-fitted model with a partial AIC of 1424.89. Since SVM performed on data reduced by logistic regression also showed promising results earlier, we ran Cox analysis on it as well to ensure a comprehensive analysis. However, the SVM model showed slightly inferior performance, achieving a concordance index of 0.79 and a partial AIC of 1615.03. Therefore, the Cox model proved to be the better-performing model for the former dataset. The reason behind it could also again be the overfitted model for SVM earlier.

Hazard ratios were also computed to quantify the relative risk of experiencing the event of interest based on different levels of the predictor variables. Hazard ratios greater than 1 indicate an increased risk, while those less than 1 suggest a reduced risk. The features with hazard ratios greater than 1 were identified as having a significant impact on the survival outcome, highlighting their importance in predicting the event occurrence(Figure 8).
Multiple cpg islands were present in the top 10 result and some of them showed their impact on liver cancer as well. For example "cg12036633" is associated with gene HERC1[18], an E3 ubiquitin ligase, which is implicated in regulating the RAF/MEK/ERK cascade[19]. In liver cancer, dysregulation of the ERK pathway can lead to increased cell proliferation, survival, and migration. Moreover, the gene associated with "cg14453201", VEGFA, which is associated with liver cancer by promoting invasion, mobilization of bone marrow-derived cells (BMDCs), and establishing a pre-metastatic niche, driving hepatic metastasis of colorectal cancer.

The clinical data underwent Kaplan-Meier curve analysis for evaluating the significance of various features concerning the disease. This was conducted to ascertain how these features could influence the prognosis and survival outcome, enabling a better understanding of their impact on the disease progression and patient outcomes. As the columns "pathology_T_stage", "residual_tumor" and "gender" displayed high hazard ratios, they were subjected to analysis. Although the "gender" column did not exhibit significant differences in the curves (Figure 9 (a)), it was observed that females had a slightly higher curve than males, suggesting a higher survival rate in female population. The Kaplan-Meier plot based on residual tumor r0,r1,r2 and rx (Figure 9 (b)), and primary tumor T stage, T1, T2, T3 and T4 (Figure 9 (c)), shows that patients with advanced stages for this cancer have the shortest survival time, suggesting the poorest prognosis. Even though the number of datapoints for both is clearly on a very low end, practical challenges in gathering data for these cases should be considered based on the severity of the disease and lower survival rates at advanced stages of cancer Patients with r0 (no visible residual tumor) in Figure 9 (b) have the highest survival rate, r2 (macroscopic residual tumor) the lowest, and rx (uncertain extent) limited data. R1 (microscopic residual tumor) shows intermediate survival, but data reliability is questionable based on the limited data. For the primary tumor curve(Figure 9 (c)), T1, T2, T3 have relatively longer survival probability than T4, with T3 slightly lower. Overall, tumor staging is vital for predicting outcomes and guiding aggressive treatment. Early intervention at lower disease levels may improve survival.

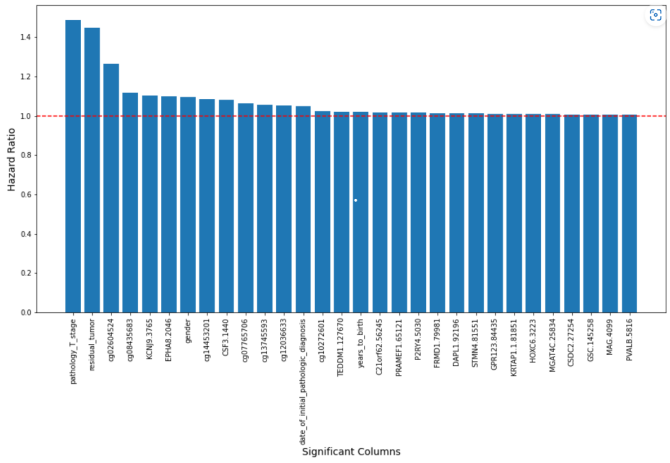

Figure 8: Top 30 Hazard Ratios for Significant Columns. Hazard ratios greater than 1 imply increased risk, while those less than 1 suggest reduced risk. Significant columns above the red dashed line (y = 1) have a significant impact on survival outcome.

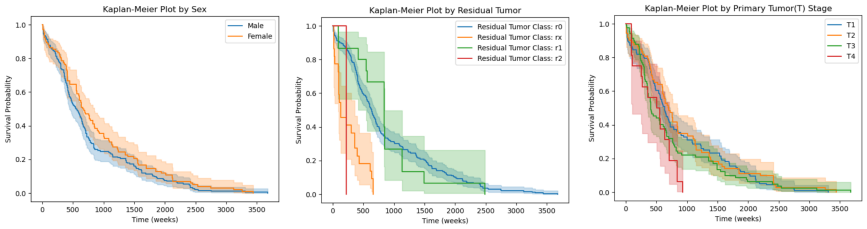

Figure 9: Kaplan-Meier plots for the features "pathology_T_stage", "residual_tumor" and "gender". The first plot on the left represents the analysis based on Sex for the column "gender". The middle plot represents the analysis of the Residual tumor, with plots for r0 (no visible residual tumor), r1 (microscopic residual tumor), r2 (macroscopic residual tumor), and rx (uncertain extent of residual tumor). The plot on the right represents the analyses of the patients based on their primary tumor T stage. T stage is denoted by a number (e.g., T1, T2, T3, T4), with higher values indicating greater local invasion and tumor size.

We conducted Weibull analysis utilizing the features identified from the overall survival analysis to predict patient survival duration, forecasting the number of days until patients might have survived. However, given the reliance on censored data, verifying result accuracy posed a challenge. To bolster the reliability of our conclusions, we implemented additional measures. Consequently, we delved into the impact of variables with high hazard ratios on the analysis. We performed a separate Weibull analysis exclusively on these high-hazard-ratio features derived from the Cox analysis. The findings exhibited a significant and notable reduction in patients' survival time, underscoring the profound statistical impact of these features on survival. For example: for patient ID TCGA-ED-A5KG-01, the number of days for survival reduced drastically from 3675 to 854 when the analysis was done using the columns with hazard ratio ¿ 1.

Discussion

The integration of multi-omics data with clinical information allowed for a comprehensive analysis of liver cancer. The results from the differential expression analysis highlighted specific genes and regions that are differentially expressed in liver cancer, confirming their potential as biomarkers. Machine learning models were successfully able to classify liver cancer samples and provide a ranked list of important features.

For the differentially expressed analysis, the scores obtained for the testing set from SVM and GB both showed low scores in accuracy, precision, sensitivity, and specificity, in spite of producing very good scores for the training data set. This suggested the overfitting of the data as discussed in the result section. Nonetheless, in spite of the overfitting of the data, the GB model was able to extract valid genes and transcripts. As mentioned before, the reason to choose GB over SVM was that it had a good average performance than SVM.

The reason it was able to extract relevant features during predictions is because the data was earlier put through differential expression analysis which reduced the features to just the highly relevant data pertaining to the cancer. Thus the machine learning algorithm was able to predict good predictions of the biologically relevant biomarkers.

For the Logistic regression-based feature reduction, in the results section, it was shown that the SVM gave excellent results, but this was not always the case. Before the transformation and scaling were done, SVM had a similar performance as GB, if not worse for the training data. After the transformation and scaling, SVM had a drastic difference in performance, but not GB. GB notably had the same performance as before, and although logistic regression feature selection could not be rerun in time, unfinished runs of it indicated that GB could have done better with a different set of features.

The drastic performance drop of training and testing for GB, accompanied by the specificity of 1 in both training and testing suggested that the overfitting was caused by balancing the data using oversampling. The minority data points, which were the negative class, were too similar to each other and thus were identified easily in comparison to the positive class.

The survival analysis conducted in this study yielded significant findings regarding liver cancer patient survivability. The Cox regression model demonstrated good predictive accuracy, identifying features with high hazard ratios that significantly impact survival outcomes. Kaplan-Meier curve analysis emphasized the importance of tumor staging and residual tumor status, revealing that patients with advanced tumor stages and visible residual tumors had shorter survival times. These results underscore the critical role of early detection and aggressive treatment in improving patient outcomes. The Weibull analysis further confirmed the influence of features with high hazard ratios on survival duration. Overall, the study highlights specific biomarkers and clinical factors that hold promise for predicting liver cancer patient survival and guiding personalized treatment approaches.

Throughout the analysis, multiple features were found which are known to be directly or indirectly associated with Liver Cancer. However, there are still some parts that could be improved. The analysis focused on binary classification and survival prediction, but future studies could explore multi-class classification for tumor stages and other survival analysis techniques. Additionally, further validation of the identified biomarkers and clinical features is essential through external independent datasets since most of the research papers which were found as proof also used the same TCGA data for analysis.

In conclusion, this project provided a comprehensive analysis of liver cancer using multi-omics data and clinical information. The integration of differential expression analysis, machine learning models, and survival analysis yielded valuable insights into liver cancer biology and potential biomarkers for diagnosis and prognosis. These findings could contribute to the development of personalized treatment strategies and improve patient outcomes in liver cancer.

Transparency

The research papers mentioned in the references were consulted while writing this project report. The help of AI like ChatGpt was used to get some insight about the topic and for paraphrasing and grammatical correction. Grammarly was also used for grammar improvement.

Author's contributions

Who of the team did what?" One paragraph per person
Blessy -Differentially expression analysis using DESeq and Limma, running machine learning algorithms (SVM and GB) on top of the data obtained from differential expression analysis, data scaling, extracting the common data in all the 4 datasets wrote data, differential expression analysis, goal of the project, abstract, data, preprocessing, results, discussion for differential expression analyis
Gabriela - Did the feature selection with logistic regression, as well as the testing and training for both classifiers with resulting data. Did the distribution data examination, transformation, and scaling. Wrote about the logistic regression, SVM, the transform and scaling part, background, and well as the result and discussion associated with the data from logistic regression feature selection.
Shubhangi - Found the initial research paper with TCGA data. Did Exploratory data analysis on the clinical data. Found about beta and m-value relation Prepared template code for data classification which involved hyperparameter tuning and cross validation, and also the evaluation which was later used in the analysis. There were other methods like random forest, xg boost and adaboost as well. Did survival analysis.

References

1. Stefano Colagrande, S.A.G.G.T.C.N.F.M. Andrea L Inghilesi: Challenges of advanced hepatocellular carcinoma. World Journal of Gastroenterology **22**(34), 7645–7659 (2016)
2. Chaudhary, K., Poirion, O.B., Lu, L., Garmire, L.X.: Deep learning–based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research **24**(6), 1248–1259 (2018)
3. Jorge A. Marrero, J.-P.B. Masatoshi Kudo: The challenge of prognosis and staging for hepatocellular carcinoma. The Oncologist **15**(4), 23–33 (2010)
4. Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., Ferrero, E., Agapow, P.-M., Zietz, M., Hoffman, M.M., *et al.*: Opportunities and obstacles for deep learning in biology and medicine. Journal of The Royal Society Interface **15**(141), 20170387 (2018)
5. Pan Du, C.-C.H.N.J.W.A.K.L.H..S.M.L. Xiao Zhang: Comparison of beta-value and m-value methods for quantifying methylation levels by microarray analysis. BMC Bioinformatics (2010)
6. Pykes, K.: Fighting Overfitting With L1 or L2 Regularization: Which One Is Better? https://neptune.ai/blog/fighting-overfitting-with-l1-or-l2-regularization (2023)
7. Jahn, K.: Data science in the life sciences - Support vector machines (2023). https://mycampus.imp.fu-berlin.de/access/content/group/c69f090a-7427-439b-ac96-40c695b75e55/Lectures/svm.pdf
8. Qi, Y.: Random forest for bioinformatics. Ensemble machine learning: Methods and applications, 307–323 (2012)
9. Song, Y.-Y., Ying, L.: Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry **27**(2), 130 (2015)
10. Lunn, M., McNeil, D.: Applying cox regression to competing risks. Biometrics, 524–532 (1995)
11. Lu, Y., Wu, M., Fu, J., Sun, Y., Furukawa, K., Ling, J., Qin, X., Chiao, P.J.: The overexpression of long intergenic ncrna00162 induced by rela/p65 promotes growth of pancreatic ductal adenocarcinoma. Cell Proliferation **53**(5), 12805 (2020)
12. Jiang, D.C.B.L.Y.K.G.H.W.G.: Ferm family proteins and their importance in cellular movements and wound healing (review) (2014)
13. Pratama, M.Y., Cavalletto, L., Tiribelli, C., Chemello, L., Pascut, D.: Selection and validation of mir-1280 as a suitable endogenous normalizer for qrt-pcr analysis of serum microrna expression in hepatocellular carcinoma. Scientific Reports **10**(1), 3128 (2020)
14. Cunningham, F., Allen, J.E., Allen, J., Alvarez-Jarreta, J., Amode, M.R., Armean, I.M., Austine-Orimoloye, O., Azov, A.G., Barnes, I., Bennett, R., Berry, A., Bhai, J., Bignell, A., Billis, K., Boddu, S., Brooks, L., Charkhchi, M., Cummins, C., Fioretto, L.D.R., Davidson, C., Dodiya, K., Donaldson, S., Houdaigui, B.E., Naboulsi, T.E., Fatima, R., Giron, C.G., Genez, T., Martinez, J.G., Guijarro-Clarke, C., Gymer, A., Hardy, M., Hollis, Z., Hourlier, T., Hunt, T., Juettemann, T., Kaikala, V., Kay, M., Lavidas, I., Le, T., Lemos, D., Marugán, J.C., Mohanan, S., Mushtaq, A., Naven, M., Ogeh, D.N., Parker, A., Parton, A., Perry, M., Pili˘zota, I., Prosovetskaia, I., Sakthivel, M.P., Salam, A.I.A., Schmitt, B.M., Schuilenburg, H., Sheppard, D., Pérez-Silva, J.G., Stark, W., Steed, E., Sutinen, K., Sukumaran, R., Sumathipala, D., Suner, M.-M., Szpak, M., Thormann, A., Tricomi, F.F., omez, D.U.-G., Veidenberg, A., Walsh, T.A., Walts, B., Willhoft, N., Winterbottom, A., Wass, E., Chakiachvili, M., Flint, B., Frankish, A., Giorgetti, S., Haggerty, L., Hunt, S.E., IIsley, G.R., Loveland, J.E., Martin, F.J., Moore, B., Mudge, J.M., Muffato, M., Perry, E., Ruffier, M., Tate, J., Thybert, D., Trevanion, S.J., Dyer, S., Harrison, P.W., Howe, K.L., Yates, A.D., Zerbino, D.R., Flicek, P.: Ensembl 2022. Nucleic Acids Res. **50**(1), 988–995 (2022). doi:10.1093/nar/gkab1049
15. Zhu, Z., Xu, J., Wu, X., Lin, S., Li, L., Ye, W., Huang, Z.: In silico identification of contradictory role of adamts5 in hepatocellular carcinoma. Technology in Cancer Research & Treatment **20**, 1533033820986826 (2021)
16. Su, J.-Q., Lai, P.-Y., Hu, P.-H., Hu, J.-M., Chang, P.-K., Chen, C.-Y., Wu, J.-J., Lin, Y.-J., Sun, C.-A., Yang, T., Hsu, C.-H., Lin, H.-C., Chou, Y.-C.: Differential DNA methylation analysis of SUMF2, ADAMTS5, and

PXDN provides novel insights into colorectal cancer prognosis prediction in taiwan. World J Gastroenterol **28**(8), 825–839 (2022). doi:10.3748/wjg.v28.i8.825
17. Martin, J., Petrillo, A., Smyth, E.C., Shaida, N., Khwaja, S., Cheow, H.K., Duckworth, A., Heister, P., Praseedom, R., Jah, A., Balakrishnan, A., Harper, S., Liau, S., Kosmoliaptsis, V., Huguet, E.: Colorectal liver metastases: Current management and future perspectives. World J Clin Oncol **11**(10), 761–808 (2020). doi:10.5306/wjco.v11.i10.761
18. Schneider, T., Martinez-Martinez, A., Cubillos-Rojas, M., Bartrons, R., Ventura, F., Rosa, J.L.: The e3 ubiquitin ligase herc1 controls the erk signaling pathway targeting c-raf for degradation. Oncotarget **9**(59), 31531 (2018)
19. Qin, Q.-F., Li, X.-J., Li, Y.-S., Zhang, W.K., Tian, G.-H., Shang, H.-C., Tang, H.-B.: Ampk-erk/carm1 signaling pathways affect autophagy of hepatic cells in samples of liver cancer patients. Frontiers in Oncology **9**, 1247 (2019)

Appendix

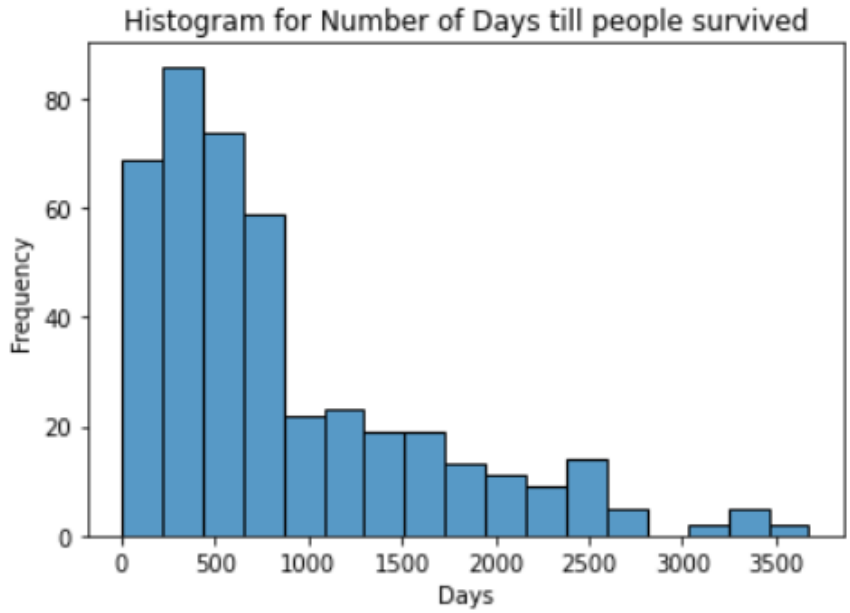

Figure 10: Histogram representing the distribution of patients who were alive at a particular number of days. The y-axis represents the number of days, while the x-axis shows the count of patients. The histogram provides insights into the survival patterns of the patient population, indicating the number of patients who survived up to specific time intervals

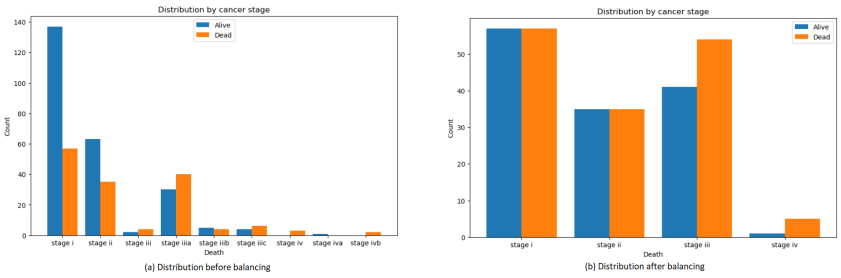

Figure 11: (a) Bar graph depicting imbalanced data for surviving and dead patients in different tumor stage groups. Stages 1 and 2 show an overrepresentation of surviving patients. (b) Bar graph showcasing balanced data after addressing the imbalances through under-sampling and merging subclasses for stages 3 and 4.

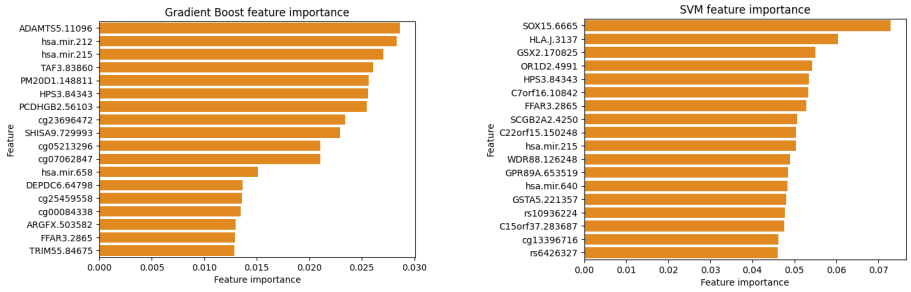

Figure 12: Top 18 feature importance list for GBestimator (left) and SVM (right).

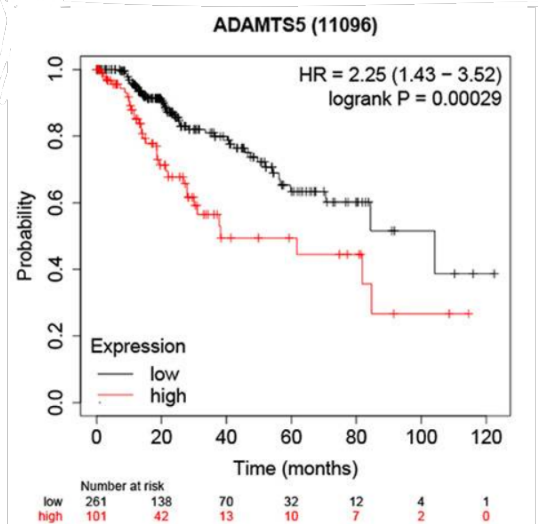

Figure 13: Kaplan-Meier plot for one of the dataset used by Zhu et. al. showing the different survival rates over time for low and high ADAMTS5 expression.