Survival Prediction for Liver Cancer Patients

Akash Agrawal, Anuj Bhardwaj, Arnav Gaur, Saksham Bathla

Department of Computer & Information Science & Engineering/Master of Science/Computer Science, University of Florida, Gainesville, Florida, USA

☯These authors contributed equally to this work.

Abstract

In 2018, in the United States alone, 1,708,921 new cancer cases were reported and almost 30% of these people died the same year. It is a group of diseases that involves abnormal cell growth which has the potential to spread to the entire body. One of the major challenges with cancer is the prognosis prediction ability. We propose a model to classify survival for liver cancer (Hepatocellular carcinoma (HCC)) patients. We identify the differentially expressed genes [2] between the sequences for tumor and normal tissue, and augment them with some clinical data variables to come up with the data set. We then use different Machine Learning techniques to classify the patients into two classes corresponding with poor (less than 1 year) and good (more than 1 year) prognosis. This can be expanded to more sophisticated studies involving different kinds of cancer [3] and more sophisticated Machine Learning models.

Author summary

Using TCGA gene expression and clinical data, we trained different classifiers for the task of prognosis prediction. We identified different clinical features using Kapan-Meir survival [11][12] analysis and identified deferentially expressed genes using voom [8] and glmfit. After feature extraction, we divided our data into 80:20 test-train split and compared the performance of various classification algorithms over the data. We were able to show that this outperforms the baseline of selecting the majority class. This pipeline can be further extended to different cancer types and based on data availability, different machine learning algorithms can be used.

Introduction

The Cancer Genome Atlas Program provided the much-needed quality genetic data for a range of cancers. This opens up a lot of avenues for using statistical and machine learning methods to perform a range of classification and prediction tasks. Providing accurate survivability and prognosis remains a sizeable challenge and we aim to help predict this based on the tumor genomic data.

Materials and methods

Data Used

The Cancer Genome Atlas (TCGA), a landmark cancer genomics program, molecularly characterized over 20,000 primary cancer and matched normal samples spanning 33

cancer types. This joint effort between NCI and the National Human Genome Research Institute began in 2006, bringing together researchers from diverse disciplines and multiple institutions.

Data used for this project was sourced from TCGA Program. We are using the clinical data and HTSeq - Counts for TGCA-LIHC (Liver cancer) to train a model to classify . This contains data for 371 primary tumors and 50 normal tissue samples. Of the total, 255 people were reported alive, whereas 164 were reported dead.

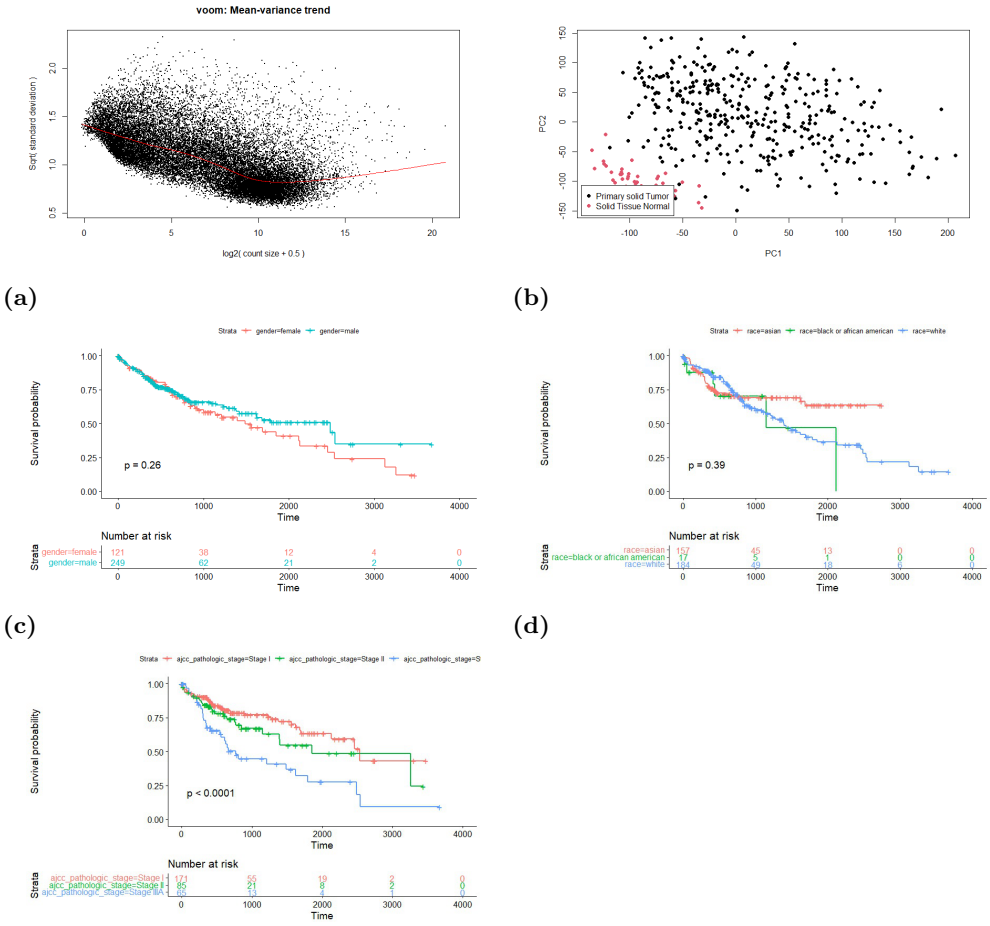


Fig 1. Primary Data Analysis **(a)** Mean-variance trend using Voom. **(b)** Dimensionality reduction analysis between healthy and tumor tissue using PCA **(c)** Kaplan-Meir survival estimation to show the difference between the prognosis of two genders. **(d)** Kaplan-Meir survival estimation to show the difference between the prognosis of people from different races. **(e)** Kaplan-Meir survival estimation to show the difference between prognosis at different clinical cancer stages.

Method Links

We used the following data for the Survival Prediction For Liver Cancer Patients:

- `https://portal.gdc.cancer.gov/projects/TCGA-LIHC`

The code for the project can be found at:

- https://github.com/sakshambathla/ml_genomics_project

Data Pre-processing and Feature Engineering

We used the TCGABiolinks package in R [1][7] to pull in the TCGA data [4]. The data is then fed through a limma pipeline, transformed using voom (Fig1), and fit over a linear model to identify the genes with differential expression. A total of 88 such genes are identified. This expression data is then augmented with the clinical data of patients to finalize the data to be used for the classification tasks with a set of 92 features for each patient.

We performed exploratory data analysis over various clinical features. We did a Kaplan-Meir survival analysis [6] to identify relevant clinical features and included them in our selected features. survival plots for gender, race, and clinical stage are given in Fig 1

From the voom plot in Fig 1(a), we can see that there is a moderate to high-level biological variation based on the voom plot.

Method Description

With feature selection complete, we used various supervised learning classification techniques to classify the patients into a high and low-risk groups.[5][8]

Linear Discriminant Analysis: Based on Fisher's linear discriminant, this model detects a linear equation of features that can be used to make a division between instances belonging to different classes.[9]

Logistic Regression: Similar to linear regression, it is also used to learn the relationship between input features and a class or label. But what logistic regression does is it predicts a probability of a data sample belonging to a particular class, using an activation function.

Decision Tree: This is an unsupervised learning technique that is used to classify the sample input into multiple classes by following a tree-like path where the decision to follow which path is based on a certain threshold for each feature. In a decision tree, every leaf node is a label.

K-Nearest Neighbor: It is supervised learning where we calculate the distance of a sample input with each data point to identify K-nearest neighbors. Based on the label of these K-nearest neighbor, we calculate the label for the sample.

Support Vector Machine: It is a supervised learning method that basically uses a hyper-plane to make a division between instances belonging to different classes.

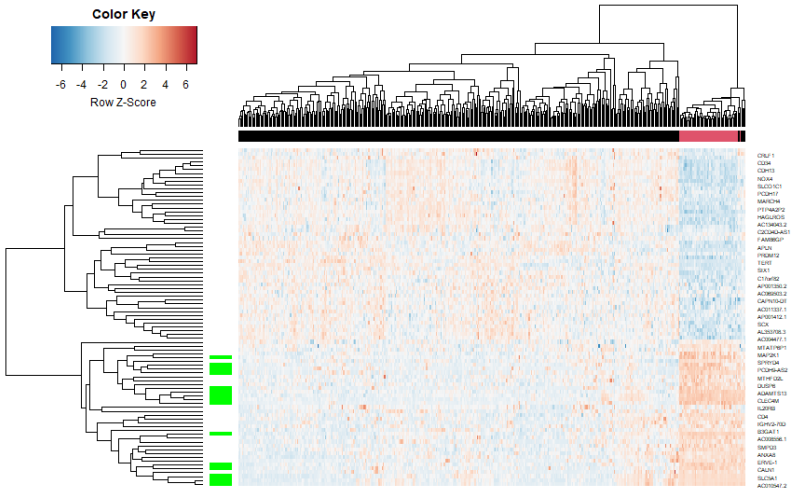
The results and comparisons of these can be seen in Fig 3

Support Methods

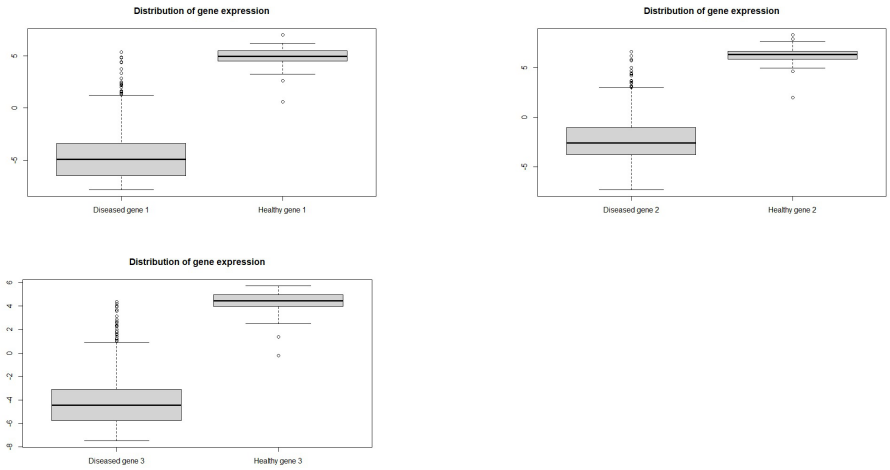
For this study, we have used Complete Linkage Clustering[13] for verifying that the genes identified in the differential gene expression analysis are correlated to the ones identified using CLC. This correlation can be seen in the heatmap in Figure 2(a). Compete linkage Clustering is one of many agglomerative hierarchical clustering algorithms. In this procedure, at each step, the two closest clusters with the shortest distance are combined into one. .

Method Issues

We faced an issue around our data pipeline. GDC released a new version on 29th March 2022 which we noticed on 4th April. They deprecated the HTSeq pipeline and replaced it with the STAR pipeline. This broke our data and feature extraction pipeline. While we had saved the final training data, the data extraction code is not working as of 7th April 2022. This hampered our ability to extend our code to different cancer types.



(a)



(b)

Fig 2. Analysis of differential gene expression a) Correlation heatmap between genes identified in GE and Complete linkage Clustering. **b)** Difference between the expression of the top 3 differently expressed genes between healthy and diseased tissue.

Method Success

We used the following metrics to evaluate the success of our model: Specificity, Accuracy, Precision, Recall,F-Score. The baseline for these models is an F1 Score of 0.64. Based on this baseline, KNN and LDA are the 2 models doing the best job.

Results

We trained 5 different classification models with an 80:20 test-train split and report the performance metrics of accuracy, precision, recall, specificity, and F1 score for all 5 models (ref. Table 1).

Summary

The baseline for these models is an F1 Score of 0.64 which is identified based on assignment to the majority class. We are able to get F1 score as high as 0.85 using KNN, which is a significant improvement over the baseline.

We notice that methods like KNN, Decision Trees and SVM have high specificity, which translates to fewer false positive results. Similarly we notice that Logistic regression is slight overfitting. Overall, LDA and KNN are the best performing models.

These results can definitely be extended to different cancers to verify if this type of pipeline would be effective in prognosis prediction for cancer patients.

Success Metrics

We constructed a confusion matrix as a unit to measure the success metric. The confusion matrix was obtained based on the performance of the models i.e. based on the four values obtained from each model: True Positives, True Negatives, False Positives, and False Negatives. Using the above-mentioned values, we calculate the Specificity, Accuracy, Precision, Recall, and F1-Score.

The F1 Score considers both precision and recall and it is the harmonic mean of the two. F1 Score is best here since there is some similarity between precision (p) and recall (r) in the system. On the contrary, if the F1 Score falls towards the lower side if one measure is improved at the expense of the other. We can confidently hold the F1-Score as the best success metric or optimization metric as we have an uneven class distribution associated with the data.

We notice that KNN has the best accuracy when compared to other models and indeed has the highest F1 score among all other models.

Issues

The major issue that we faced with regards to the results was the fact that the number of patients is not very large and most of the models we trained were highly overfitted and didn't perform particularly on testing data. To overcome this, we had to choose algorithms that work satisfactorily in the case of small amounts of data as well. We also skipped extensive hyper-parameter tuning to avoid overfitting issues.

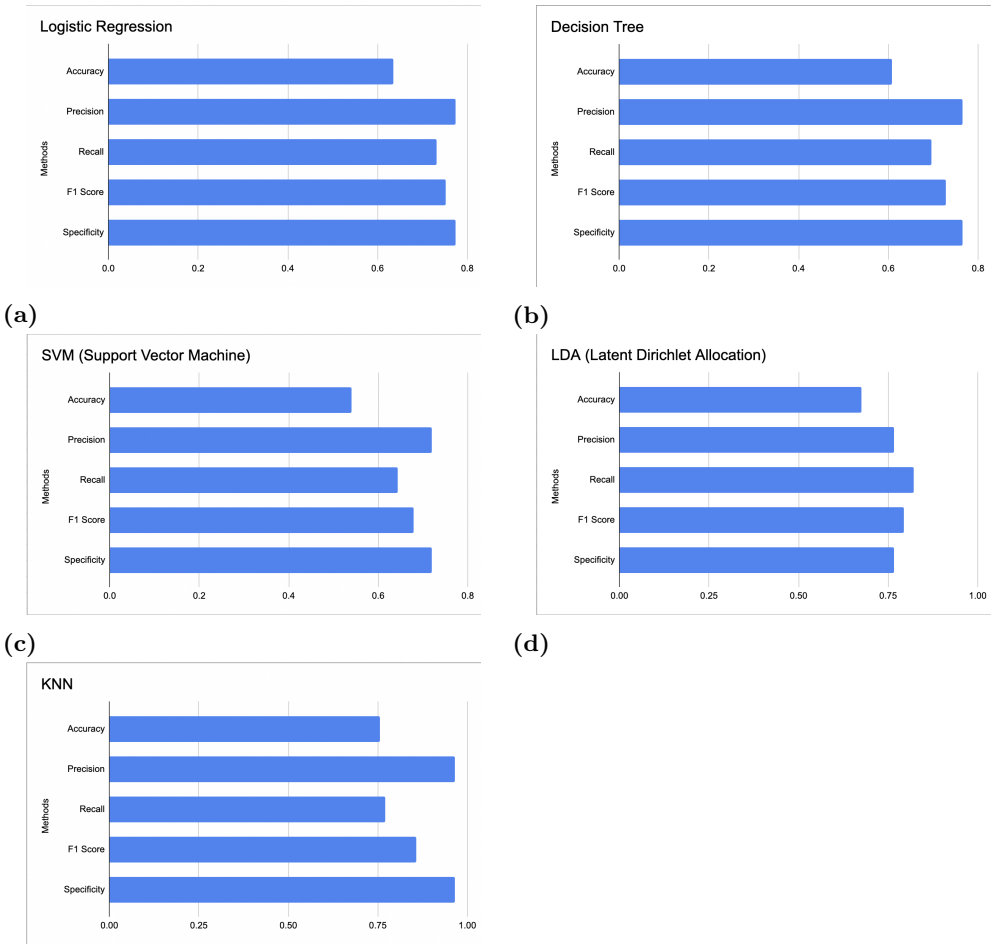


Fig 3. Primary Data Analysis (a) Mean-variance trend using Voom. **(b)** Dimensionality reduction analysis between healthy and tumor tissue using PCA **(c)** Kaplan Meir survival estimation to show the differences between prognosis of two genders. **(d)** Kaplan Meir survival estimation to show the differences between prognosis of three race. **(e)** Kaplan Meir survival estimation to show the differences between prognosis at different stages.

Table 1. Metrics for performance of Logistic Regression with the Confusion Matrix

Methods	Logistic Regression
Accuracy	0.6351
Precision	0.7735
Recall	0.7321
*F*1*score*	0.7522
Specificity	0.7735

	True +ve	True -ve
Predicted +ve	41	15
Predicted -ve	12	6

Table 2. Metrics for performance of Decision Trees with the Confusion Matrix

Methods	Decision Trees
Accuracy	0.6081
Precision	0.7647
Recall	0.6964
*F*1*score*	0.7289
Specificity	0.7647

	True +ve	True -ve
Predicted +ve	39	17
Predicted -ve	12	6

Table 3. Metrics for performance of Support Vector Machine with the Confusion Matrix

Methods	SVM
Accuracy	0.5405
Precision	0.7200
Recall	0.6428
*F*1*score*	0.6792
Specificity	0.7200

	True +ve	True -ve
Predicted +ve	36	20
Predicted -ve	14	4

Table 4. Metrics for performance of Latent Dirichlet allocation with the Confusion Matrix

Methods	LDA
Accuracy	0.6756
Precision	0.7666
Recall	0.8214
*F*1*score*	0.7931
Specificity	0.7666

	True +ve	True -ve
Predicted +ve	46	10
Predicted -ve	14	4

Table 5. Metrics for performance of K-nearest neighbour with the Confusion Matrix

Methods	KNN
Accuracy	0.7567
Precision	0.9642
Recall	0.7714
*F*1*score*	0.8571
Specificity	0.9642

	True +ve	True -ve
Predicted +ve	54	16
Predicted -ve	2	2

Conclusion

The purpose of this study was to classify survival for liver cancer (Hepatocellular carcinoma (HCC)) patients into high-risk and low-risk patients. We trained 5 different models and identified KNN to have the highest accuracy while making this classification and we also identified these models to perform better than the baseline. We can conclude that using gene expression data in addition to some clinical features can be a fairly effective in predicting prognosis in liver cancer patients based on the genomics profile of cancerous tissue.

Given the recent increase in the genomics data availability, machine learning can be used to do prognosis prediction as demonstrated in this study. This approach can be extended to different kinds of cancer as well as many other diseases.

Discussion

Prognosis prediction is one of the most researched field in medical science. With the advent of widely available genomic data, there are new avenues open for using data analysis and machine learning techniques in prognosis prediction.

Based on clinical and genomic data, we identified relevant genes and features and compared different classification models for the purpose of prognosis prediction.

We were able to get significantly better than baseline performance while identifying low and high risk patients. While we faced some issues due to GDC updates, rending our pipeline useless, we were still able to successfully demonstrate that a combination of clinical and genomics data is the way to go.

Novelty: There has been a significant amount of research in data analysis of cancer patient survival but not a lot of those used machine learning techniques. With the recent uptick in the amount of genomics data available, using Machine Learning for prognosis prediction, is much better option.

References

1. Matthew E. Ritchie, Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, Gordon K. Smyth. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. 2015 Apr 20; 43(7): e47. https://academic.oup.com/nar/article/43/7/e47/2414268?login=false

2. Christine Steinhoff, Martin Vingron. Normalization and quantification of differential expression in gene expression microarrays. Briefings in Bioinformatics, Volume 7, Issue 2, June 2006, Pages 166–177 , https://doi.org/10.1093/bib/bbl002.

3. The ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Pan-cancer analysis of whole genomes. Nature 578, 82–93 (2020). https://doi.org/10.1038/s41586-020-1969-6

4. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge Published online 2015 Jan 20. doi: 10.5114/wo.2014.47136 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4322527/

5. A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data Published: August 13, 2014 https://doi.org/10.1371/journal.pone.0103207

6. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics Volume 173, Issue 2, 5 April 2018, Pages 400-416.e11 https://doi.org/10.1016/j.cell.2018.02.052

7. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data Nucleic Acids Research, Volume 44, Issue 8, 5 May 2016, Page e71 https://doi.org/10.1093/nar/gkv1507

8. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts 2014 Feb 3;15(2):R29. doi: 10.1186/gb-2014-15-2-r29. https://pubmed.ncbi.nlm.nih.gov/24485249/

9. The design and implementation of VOOM: a parallel virtual Object Oriented machine Microprocessing and Microprogramming Volume 32, Issues 1–5, August 1991, Pages 289-296 https://doi.org/10.1016/0165-6074(91)90360-6

10. limma: Linear Models for Microarray Data Part of the Statistics for Biology and Health book series (SBH) Bioinformatics and Computational Biology Solutions Using R and Bioconductor pp 397-420— https://link.springer.com/chapter/10.1007/0-387-29362-$0_2$3

11. Understanding survival analysis: Kaplan-Meier estimate [11] Int J Ayurveda Res. 2010 Oct-Dec; 1(4): 274–278. doi: 10.4103/0974-7788.76794 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3059453/

12. A Practical Guide To Understanding Kaplan-Meier Curves [12] Otolaryngol Head Neck Surg. 2010 Sep; 143(3): 331–336. doi: 10.1016/j.otohns.2010.05.007articles/PMC3059453/ https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3932959/

13. Hierarchical cluster analysis in clinical research with heterogeneous study population: highlighting its visualization with R Ann Transl Med. 2017;5(4):75. doi:10.21037/atm.2017.02.05 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5337204/