Utilizing Machine Learning Techniques for Cancer Type and Cancer Stage Classification Based on Gene Expression Data of Hepatobiliary Cancer

Michael Zheng, Shivank Sadasivan

1. ABSTRACT

This study addresses the critical need for precise differentiation between hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA), the two major subtypes of liver cancer, by leveraging advanced machine learning techniques and gene expression data from the cBioPortal Pan-Cancer study. We employed principal component analysis (PCA) to visualize and confirm the distinctiveness of HCC and CCA at the molecular level. Additionally, several machine learning classifiers, including Logistic Regression, Random Forest, K-Nearest Neighbours (K-NN), and Support Vector Machines (SVM), were implemented to classify the cancer subtypes and stages. The models were evaluated using a 5-fold cross-validation approach, and their performance was quantified through metrics such as Accuracy, F1 Score, and Precision. While our results confirmed the effectiveness of gene expression profiling in distinguishing between HCC and CCA, challenges in accurate stage classification of HCC highlighted potential areas for further enhancement of the models for precision medicine applications. This study thus provides a foundation for improving diagnostic strategies and treatment decision-making in liver cancer care.

2. INTRODUCTION

Liver cancer, one of the leading causes of cancer-related deaths worldwide, primarily manifests in two distinct subtypes: hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA). These subtypes differ significantly in their pathology, prognosis, and treatment responses. HCC, the most common form, originates in hepatocytes, while CCA arises from the bile duct cells within the liver. Studying these subtypes is crucial as they can coexist in a condition known as combined hepatocellular-cholangiocarcinoma, posing diagnostic and therapeutic challenges. [1] [2]

The objective of this study is to leverage the vast array of genomic data available through the cBioPortal within the Pan-Cancer Atlas initiative. This initiative provides access to a comprehensive dataset, including gene expression profiles from liver cancer patients, standardized to accommodate variations across different sequencing platforms and batches. This normalization ensures robust comparative analyses across diverse patient samples.

3. METHODS

I. Data Acquisition and Pre-processing

The primary dataset comprises gene expression profiles integrated with demographic and clinical information from 2970 patients (Data Set 1), which includes details such as cancer type, stage, grade, and unique sample identifiers. This integration (Metadata Integration) enhances the robustness of our analyses by ensuring each clinical profile correlates with specific gene expression data, facilitating accurate disease characterization.

Further pre-processing steps involved using the normalization of gene expression data (Data Set 2), crucial for consistent comparisons across the dataset. Discrepancies in cancer staging notation were standardized to numerical values in a process we termed 'Stage Data Harmonization', simplifying classification tasks and streamlining subsequent machine learning analyses.
The selection of the liver cancer cohort for this study was strategically made to ensure a balanced representation of early and late-stage cases, providing a comprehensive view of the disease's progression and variance across stages. This selection is pivotal in addressing the complexities associated with liver cancer's heterogeneous nature.

By methodically mapping and integrating clinical data with corresponding gene expression profiles, this study sets the stage for a detailed examination of the molecular distinctions between HCC and CCA, with the ultimate goal of enhancing diagnostic precision and therapeutic efficacy in liver cancer treatment.

II. Principal Component Analysis (PCA)

PCA is a critical statistical tool used in genomics to simplify the complexity in high-dimensional data by reducing its dimensions without significant loss of information. This method involves extracting eigenvalues and eigenvectors from the covariance matrix of the data, which represent the directions of maximum variance and help in identifying the principal components that capture the most significant information. The effectiveness of PCA in revealing hidden patterns and essential features in expression data makes it a cornerstone technique in bioinformatics, especially useful for large datasets where traditional methods are too time-consuming [3]

III. Logistic Regression

Logistic Regression is a fundamental statistical model used primarily for binary classification tasks. In the context of this study, it was employed to differentiate between the subtypes and stages of liver cancer based on gene expression data. The model operates by estimating probabilities using a logistic function, which is particularly effective in cases where the response variable is

categorical. Logistic Regression is valued for its simplicity and interpretability, as it provides direct insight into the influence of each feature on the probability of belonging to a particular class. This makes it a suitable choice for medical applications where understanding the impact of variables is crucial. [4]

IV. Random Forest

Random Forest is an ensemble learning method known for its robustness and accuracy. It constructs multiple decision trees during training and outputs the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random Forest handles large datasets with high dimensionality well, making it ideal for gene expression data in cancer research. It provides not only high predictive accuracy but also measures of feature importance, which can be crucial for identifying key genetic markers relevant to liver cancer subtypes and stages. The inherent diversity in the ensemble reduces the risk of overfitting, enhancing the model's performance on unseen data.[5]

V. K-Nearest Neighbours (K-NN)

K-NN is a non-parametric method used for classification and regression. For our study, K-NN was applied to classify liver cancer stages and subtypes by identifying the most common class among the k-nearest samples. This method is intuitive and simple, relying on the assumption that similar things exist in close proximity. K-NN can be particularly effective in small datasets, as it makes decisions based on the localized landscape of the data, which is essential when dealing with limited samples. The choice of 'k' and the distance metric significantly influence its effectiveness and were carefully tuned based on our dataset characteristics. [4]

VI. Support Vector Machines (SVM)

SVM are powerful supervised learning models used for classification and regression challenges. SVM works by finding the hyperplane that best divides a dataset into classes, which is particularly useful for distinguishing between different cancer subtypes and stages in our study. The strength of SVM lies in its ability to handle non-linear data using kernel tricks, enabling it to model complex relationships between gene expression patterns and cancer characteristics. SVM is noted for its effectiveness in high-dimensional spaces, such as gene expression data, making it a valuable tool for precise and robust cancer classification. [4]

4. RESULTS

The initial analytical technique applied to the pre-processed dataset was Principal Component Analysis (PCA). We conducted PCA manually, following these steps:

- Covariance Matrix Computation: Calculate the covariance matrix to understand how variables in the dataset vary with each other.
- Eigen decomposition: Perform eigen decomposition on the covariance matrix to extract the eigenvalues and eigenvectors.
- Selection of Principal Components: Choose the two eigenvectors with the largest eigenvalues for simplicity in visualization and interpretation.

PCA is crucial as it reveals the principal components where the variability in the dataset is most significant, essential for high-dimensional data such as gene expression profiles. By reducing dimensions, PCA helps in identifying underlying relationships within the data, indicating viable paths for further modelling.

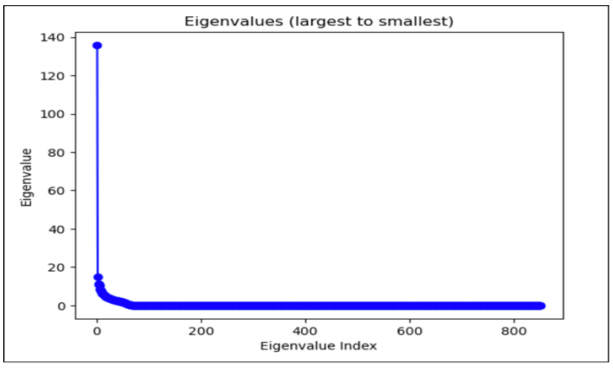

Figure 1: Eigenvalues ordered from largest to smallest

Following information was obtained from the above figure by performing eigen decomposition on the covariance matrix and obtaining the eigenvalues. The main takeaway from the above figure is that most of the variability is captured using only a few features. This is important because when we utilize other models such as logistic regression, we can remove many features and simplify our model without losing much information from the dataset.

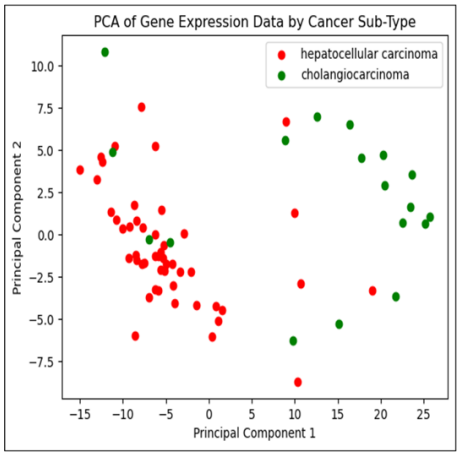

Figure 2: Relationships in the dataset for cancer subtypes after being projected to 2 dimensions

Figure 2 illustrates the projection of our dataset into two dimensions using the first two principal components (PC1 and PC2), derived from the eigenvectors with the largest eigenvalues. This two-dimensional PCA plot reveals distinct clustering of gene expression profiles corresponding to hepatocellular carcinoma (HCC) and cholangiocarcinoma (CCA). The clear differentiation between these subtypes in the PCA plot justifies the application of classification models such as K-Nearest Neighbours and Logistic Regression, with an anticipation of effective performance due to the discernible differences in gene expression patterns.

- **PCA of Cholangiocarcinoma stages**

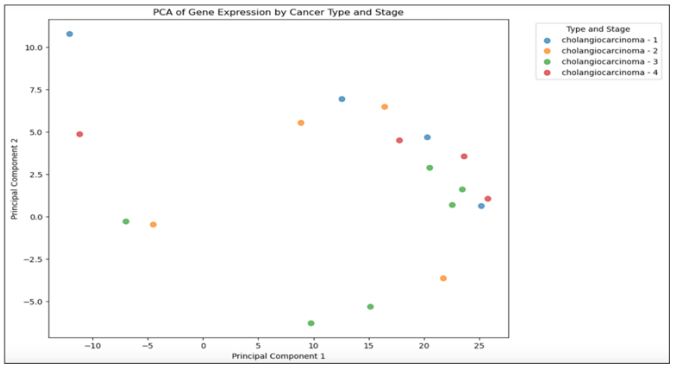

Figure 3: PCA of Gene Expression for Stages of Cholangiocarcinoma

- **PCA of Hepatocellular Carcinoma stages**

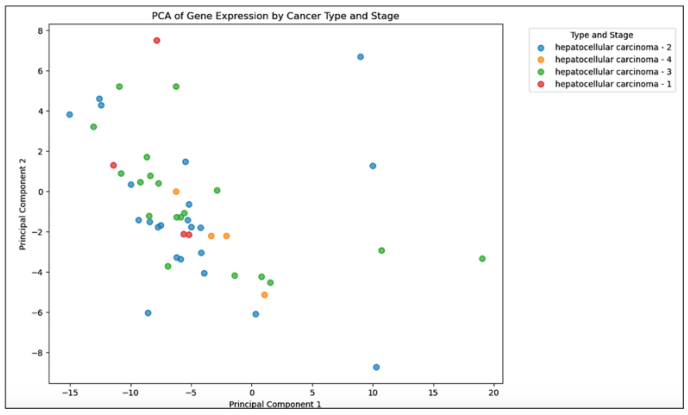

Figure 4: PCA of Hepatocellular Carcinoma by Stages

Similar to Figure 2, Figures 3 and 4 employ PCA to investigate the relationship between different stages of CCA and HCC, respectively. These visualizations indicate a correlation between gene expression and cancer stages, although the relationships appear less distinct compared to the subtype analysis. The murkier differentiation in stages suggests potential challenges in achieving high accuracy with stage classification models. Additionally, the disparity in sample sizes between HCC and CCA, particularly the smaller sample set for CCA, led us to primarily focus our subsequent analysis and model tuning on HCC.

After implementing Principal Component Analysis, we applied the K-Nearest Neighbours (KNN) algorithm to further analyze our dataset. The steps involved in the KNN algorithm were as follows:

- Distance Calculation: Each data point's distance to every other point in the dataset was computed, facilitating a comparison based on proximity.
- Sorting and Neighbour Selection: All points were sorted based on their distances, and the 'k' nearest ones were selected for each point.
- Classification: Each point was classified according to the majority label among its k-nearest neighbours, leveraging the local label consensus to determine its class.

The choice of KNN was motivated by its simplicity and effectiveness, making it particularly suitable for our dataset, which consisted of 67 samples. Through trials with various values of 'k', we identified that a setting of k=2 yielded optimal classification performance. The inherent simplicity of KNN not only makes it easily tuneable but also enhances its interpretability, proving invaluable in preliminary data analysis and providing a solid foundation for subsequent analytical steps.

The Logistic Regression model, implemented using specialized packages for enhanced performance and reliability, demonstrated exceptional effectiveness in classifying liver cancer subtypes. For hepatocellular carcinoma (HCC), the model achieved perfect scores across all metrics: test accuracy, F1-score, and precision, all recorded at 1.00, with a cross-validation (CV) accuracy of 0.85. Similarly, for cholangiocarcinoma (CCA), the model also reached a perfect score in test accuracy, F1-score, and precision, clearly indicating its robustness in distinguishing between these subtypes.

Random Forest, another model applied via established statistical packages, mirrored the high performance of Logistic Regression. For HCC, the model reached a test accuracy, F1-score, and precision of 1.00, with a slightly lower CV-accuracy of 0.83 compared to Logistic Regression. The results for CCA were also perfect in terms of test accuracy, F1-score, and precision, affirming the Random Forest's capability to handle the complex patterns inherent in gene expression data effectively.

In contrast to the perfect scores of the logistic regression and random forest models, K-NN showed slightly lower metrics, particularly in CV-accuracy and precision for CCA, which were 0.75 and 0.75, respectively. For HCC, K-NN achieved a test accuracy of 0.89, an F1-score of 0.95, and perfect precision, indicating a strong but slightly less consistent performance than the other models.

Methods	Cancer Type	Test Accuracy	F1- Score	Precision	CV-Accuracy
Logistic Regression	HCC	1	1	1	0.85
	CCA		1	1	
Random Forest	HCC	1	1	1	0.83
	CCA		1	1	
K-NN	HCC	0.89	0.95	1	0.75
	CCA		0.86	0.75	

Following the initial success in differentiating liver cancer subtypes, our focus shifted to a more granular analysis of HCC stages, prompted by the richer dataset available compared to CCA. Employing PCA, we assessed the spread of HCC across various stages, consolidating stages 1 and 2 into an "Early stage" category and designating stage 3 as "Late stage". Stage 4 was excluded from the analysis to avoid the confounding effects of metastasis, which could introduce noise into the predictive modelling. In the stage-specific analysis using Logistic Regression, the model achieved a test accuracy of 0.78 for Early stage HCC, with an F1-score of 0.80 and precision of 0.67, while the cross-validation (CV) accuracy was 0.61. For Late stage HCC, the model demonstrated a test accuracy of 0.75 and a perfect F1-score of 1.00. These results suggest that Logistic Regression is moderately effective in distinguishing between early and late stages of HCC, with better performance observed in the more advanced stage.

Random Forest, known for its effectiveness in handling complex classification problems, showed varied results. The Early stage classification achieved a test accuracy of 0.56, with an F1-score of 0.67 and a precision of 0.50, alongside a CV-accuracy of 0.61. However, the Late stage performance was notably lower, with a test accuracy of just 0.33, still achieving a perfect F1-score of 1.00. These outcomes indicate challenges in the Random Forest model's ability to generalize across different stages of HCC, particularly in the later stage. The SVM model offered a balance of performance across the two stages. For the Early stage, it recorded a test accuracy of 0.67, an F1-score of 0.67, and a precision of 0.60, with a CV-accuracy of 0.51. The Late stage results were comparable, with a test accuracy of 0.67 and an F1-score of 0.75. These metrics suggest that SVM provides a consistent but modest performance across both stages.

Methods	Cancer Stage	Test Accuracy	F1- Score	Precision	CV-Accuracy
Logistic Regression	Early	0.78	0.80	0.67	0.61
	Late		0.75	1	
Random Forest	Early	0.56	0.67	0.50	0.61
	Late		0.33	1	
SVM	Early	0.67	0.67	0.60	0.51
	Late		0.67	0.75	

5. DISCUSSION

The application of machine learning models in the classification of hepatocellular carcinoma (HCC) stages has highlighted several critical insights and challenges. While the PCA effectively distinguished between HCC and cholangiocarcinoma (CCA) samples, indicating the viability of gene expression profiling for subtype differentiation, the same success was not uniformly observed across the staging of HCC. Logistic Regression and SVM provided reasonable effectiveness, but the variability in Random Forest's performance suggests that these models may require further tuning and enhancement. This differential performance underscores the complexity of stage-specific cancer classification and the need for model selection that aligns closely with the specific characteristics of the data.

Challenges were particularly noted in the classification of HCC into early and late stages. The less robust performance in early-stage classification reflects the nuanced complexity of early-stage cancer gene expression, which may not be fully captured by the current models or available data. This aspect is crucial as it directly impacts the development of targeted therapeutic strategies and personalized medicine approaches in treating HCC. The inherent variability in early-stage gene expression and potential overlap of gene expression patterns across early cancer stages, compounded by the presence of confounding clinical factors not accounted for in the analysis, further complicate the staging accuracy.

6. CONCLUSION

The study's findings point towards the potential and limitations of using machine learning models for the precise classification of liver cancer subtypes and stages. Moving forward, enriching the dataset with a broader range of demographic and genetic backgrounds could enhance the robustness and applicability of these models. Expanding the classification to include additional cancer subtypes and incorporating varied tumor characteristics such as primary versus metastatic status are essential steps to improve gene expression profiles and classification accuracy.

Moreover, future efforts should focus on addressing the identified limitations by integrating larger and more diverse datasets, which could provide a more comprehensive understanding of the disease mechanisms at play. Exploring these avenues will undoubtedly aid in the refinement of current models and potentially lead to more effective diagnostic tools and treatment options for hepatobiliary cancer

7. REFERENCES:

[1] Yu-Zhu Zhang, Yu-Chen Liu, Tong Su, Jiang-Nan Shi, Yi Huang, Bo Liang, Current advances and future directions in combined hepatocellular and cholangiocarcinoma, *Gastroenterology Report*, Volume 12, 2024, goae031, https://doi.org/10.1093/gastro/goae031

[2] Liangtao Ye, Julia S. Schneider, Najib Ben Khaled, Peter Schirmacher, Carolin Seifert, Lea Frey, Yulong He, Andreas Geier, Enrico N. De Toni, Changhua Zhang, Florian P. Reiter; Combined Hepatocellular-Cholangiocarcinoma: Biology, Diagnosis, and Management. *Liver Cancer* 9 February 2024; 13 (1): 6–28. https://doi.org/10.1159/000530700

[3] Florian Privé, Keurcien Luu, Michael G B Blum, John J McGrath, Bjarni J Vilhjálmsson, Efficient toolkit implementing best practices for principal component analysis of population genetic data, *Bioinformatics*, Volume 36, Issue 16, August 2020, Pages 4449–4457, https://doi.org/10.1093/bioinformatics/btaa520

[4] Su, F., Chao, J., Liu, P. *et al.* Prognostic models for breast cancer: based on logistics regression and Hybrid Bayesian Network. *BMC Med Inform Decis Mak* **23**, 120 (2023). https://doi.org/10.1186/s12911-023-02224-1

[5] I. Taraniya, Bhaskar Reddy P.V., Yeddula Divyasri, Varshini Chaithra, Nalam Lakshmi Raviteja; Machine learning based breast cancer detection using logistic regression. *AIP Conf. Proc.* 13 February 2024; 2742 (1): 020084. https://doi.org/10.1063/5.0200498

[6] Bostanci, E., Kocak, E., Unal, M., Guzel, M. S., Acici, K., & Asuroglu, T. (2023, March 13). *Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer*. MDPI. https://www.mdpi.com/1424-8220/23/6/3080

[7] Yu, X., Cao, S., Zhou, Y., Yu, Z., & Xu, Y. (2020, June 30). *Co-expression based cancer staging and application*. Nature News. https://www.nature.com/articles/s41598-020-67476-7

[8] Alharbi F, Vakanski A. Machine Learning Methods for Cancer Classification Using Gene Expression Data: A Review. Bioengineering (Basel). 2023 Jan 28;10(2):173. doi: 10.3390/bioengineering10020173. PMID: 36829667; PMCID:PMC9952758